

**باسمه تعالی**



**دانشگاه صنعتی مالک‌اشتر**

مجتمع دانشگاهی برق و کامپیوتر

**پايان‌نامه دوره کارشناسي ارشد رشته مهندسي کامپیوتر**

عنوان

بهبود تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای مبتني بر يادگيري ماشين

توسط: اميرحسين قاسمي

استاد راهنما: آقاي دكتر علي هادوي

استاد (اساتید) مشاور

1 خرداد ماه 1403

**باسمه تعالی**



**دانشگاه صنعتی مالک‌اشتر**

مجتمع دانشگاهی برق و کامپیوتر

**پایان­نامه­ دوره کارشناسی ارشد رشته­ی مهندسی کامپیوتر گرایش رايانش امن**

**با عنوان**

بهبود تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای مبتني بر يادگيري ماشين

**در تاریخ -/-/- توسط کمیته­ی تخصصی زیر، موردبررسی قرار گرفت و با نمره­ی ................... و درجه­ی ................... به تصویب رسید.**

|  |  |  |  |
| --- | --- | --- | --- |
| ردیف | مسئولیت | عنوان | امضای |
| **1** | **استاد راهنمای اول** | **دکتر** **علی هادوی** | ……………. |
| **2** | **استاد راهنمای دوم** | **دکتر** ……………. | ……………. |
| **3** | **استاد مشاور** | **دکتر محمدعلي كيوان‏راد** | ……………. |
| **4** | **استاد داور بیرونی** | **دکتر** ……………. | ……………. |
| **5** | **استاد داور بیرونی** | **دکتر** ……………. | ……………. |
| **6** | **استاد داور داخلی** | **دکتر** ……………. | ……………. |
| **7** | **استاد داور داخلی** | **دکتر** ……………. | ……………. |
| **8** | **مدیر تحصیلات تکمیلی مجتمع** | **دکتر** ……………. | ……………. |

تعیین سطح طبقه‌بندی پایان‌نامه

عنوان پایان‌نامه: بهبود تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای مبتني بر يادگيري ماشين

نام و نام خانوادگی استاد راهنما: **دکتر علي هادوي**

|  |  |  |
| --- | --- | --- |
| نام و نام خانوادگی دانشجو: اميرحسين قاسمي | | شماره دانشجویی: 4011417038 |
| رشته: **مهندسی كامپيوتر** | | گرایش: رايانش امن |
| ازنظر استاد راهنما این پایان‌نامه: | 1- طبقه‌بندی دارد□ | 2- طبقه‌بندی ندارد□ |
|  | | نام و امضای استاد راهنما: |

"تبصره: باتوجه‌به طبقه دار بودن پایان‌نامه، تا تاریخ ................ این پایان‌نامه، فقط با نظر این واحد دانشگاهی، در اختیار متقاضی قرار گیرد."

**تمام حقوق مادی و معنوی مترتب بر نتایج مطالعات،**

**ابتکارات و نوآوری­های حاصل از این تحقیق، متعلق**

**به دانشگاه صنعتی مالک اشتر است.**

|  |  |  |
| --- | --- | --- |
| ریاست گروه علمی | مدیر آموزش و تحصیلات تکمیلی مجتمع | ریاست مجتمع |
|  | **دکتر کوروش داداش تبار احمدی** | **دکتر علی جبار رشیدی** |

**تأیید صحت و اصالت نتایج پایان‌نامه**

اینجانب اميرحسين قاسمي به شماره دانشجویی 4011417038 دانشجوی رشته مهندسی كامپيوتر گرایش رايانش امن مقطع تحصیلی کارشناسی‌ارشد تأیید می‌نمایم که کلیه نتایج این پایان‌نامه حاصل کار اینجانب و بدون دخل‌وتصرف است و موارد نسخه‌برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع آورده‌ام.

در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با این‌جانب رفتار خواهد شد و حق هرگونه اعتراض در خصوص احقاق حقوق مکتسب و تعیین تخلف و مجازات را از خویش صلب می‌نمایم. در ضمن، مسئولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده این‌جانب خواهد بود و دانشگاه هیچ‌گونه مسئولیتی دراین‌خصوص نخواهد داشت.

فهرست مطالب

عنوان صفحه

[فهرست مطالب ‌ه](#_Toc189944715)

[فهرست جدول‌ها ‌و](#_Toc189944716)

[فهرست شکل‌ها ‌ز](#_Toc189944717)

[فهرست نمادها ‌ح](#_Toc189944718)

[کوته‌نوشت‌ها ‌ط](#_Toc189944719)

[چکيده ‌ي](#_Toc189944720)

[فصل 1- مقدمه 1](#_Toc189944721)

[1-1- بيان مسئله 2](#_Toc189944722)

[1-1-1- مسئله و اهداف اصلي تحقيق: 2](#_Toc189944723)

[1-2- تشريح و بيان موضوع: 3](#_Toc189944724)

[1-2-1- تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای 4](#_Toc189944725)

[1-2-2- بررسي علل نياز به مجموعه‌داده‏هاي قابل‌اعتماد و قوي 7](#_Toc189944726)

[1-2-3- بررسي اهميت پايگاه‏داده‏هاي غيررابطه‏اي 8](#_Toc189944727)

[1-2-4- ارائه نمونه‌ای حمله تزريق پایگاه‌داده غیررابطه‌ای 8](#_Toc189944728)

[1-2-5- بررسي و تشريح مدل 11](#_Toc189944729)

[1-3- ضرورت انجام تحقيق: 13](#_Toc189944730)

[1-3-1- بر مبناي تمايل بازار كار: 15](#_Toc189944731)

[1-3-2- بر مبناي نياز به توسعه سريع: 15](#_Toc189944732)

[1-3-3- بر مبناي تقاضاي جهاني و داخلي: 16](#_Toc189944733)

[1-3-4- بر مبناي اهميت هوش مصنوعي 16](#_Toc189944734)

[1-3-5- بر مبناي اهميت حملات تزريق 16](#_Toc189944735)

[1-4- سابقه تحقيقات و مطالعات انجام‌گرفته 17](#_Toc189944736)

[1-5- هدف پژوهش 21](#_Toc189944737)

[1-6- نوآوري پژوهش 23](#_Toc189944738)

[1-7- ساختار فصول پایان‌نامه 25](#_Toc189944739)

[فصل 2- مبانی نظری و پیشینه پژوهش 27](#_Toc189944740)

[2-1 مبانی نظری 28](#_Toc189944741)

[2-1-1- شناخت پایگاه‌داده‌های غیررابطه‌ای 28](#_Toc189944742)

[2-1-2- بررسی نقاط ضعف و چالش‌های امنیتی مرتبط با حملات تزریق داده در محیط‌های پایگاه‌داده غیررابطه‌ای 29](#_Toc189944743)

[2-1-3- یادگیری ماشین 34](#_Toc189944744)

[2-1-4- معرفی مدل‌های زبان بزرگ و مقایسه آن‌ها با روش‌های سنتی درتولید داده‌های مصنوعی 38](#_Toc189944745)

[2-2- تزریق پایگاه‌داده غیررابطه‌ای 41](#_Toc189944746)

[2-2-1- رویکرد یادگیری ماشین 53](#_Toc189944747)

[2-2-2- رویکرد غیر یادگیری ماشین 55](#_Toc189944748)

[فصل 3- روش پیشنهادی 59](#_Toc189944749)

[**3-2** **معرفی و توصیف رویکرد کلی پژوهش** 60](#_Toc189944750)

[3-3 روش جمع‌آوری داده 63](#_Toc189944751)

[**3-3-1 استخراج ویژگی‌های کلیدی از مجموعه‌داده اولیه** 63](#_Toc189944752)

[**3-3-2** **تولید داده‌های مصنوعی با استفاده از مدل GPT** 65](#_Toc189944753)

[**3-4** معرفی مدل‌های Logistic Regression، Random Forest، و XGBoost، 66](#_Toc189944754)

[**3-5 ارائه جزئیات مربوط به تنظیم پارامترها** 68](#_Toc189944755)

[3-5 پیاده‌سازی 73](#_Toc189944756)

[فصل 4- تجزیه‌وتحلیل داده‌ها (یافته‌ها) 79](#_Toc189944757)

[**4-1 معیارهای ارزیابی عملکرد** 80](#_Toc189944758)

[4-1-1- - خطای میانگین مربعات 80](#_Toc189944759)

[4-1-2- میانگین خطای مطلق 81](#_Toc189944760)

[4-1-3- میانگین درصد خطای مطلق 81](#_Toc189944761)

[4-1-4- صحت 82](#_Toc189944762)

[4-1-5 ناحیه زیر منحنی 82](#_Toc189944763)

[4-1-6 حساسیت 82](#_Toc189944764)

[4-1-7 ویژگی Specificity 83](#_Toc189944765)

[4-1-8 دقت پیش‌بینی 83](#_Toc189944766)

[4-1-9 تحلیل معیارها در پروژه ما 83](#_Toc189944767)

[**4-2** **بررسی دیتاست‌ها** 84](#_Toc189944768)

[**4-3 مقایسه عملکرد مدل‌ها** 90](#_Toc189944769)

[**4-3-1 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست اصلی** 90](#_Toc189944770)

[**4-3-2 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی** 96](#_Toc189944771)

[**4-3-3 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی به عنوان داده اموزشی و دادهای اصلی به عنوان داده آزمایشی** 102](#_Toc189944772)

[فصل 5: بحث، نتیجه گیري و پیشنهادات 109](#_Toc189944773)

[**5-1 مرور اجمالی نتایج تحقیق** 110](#_Toc189944774)

[**5-2 تحلیل نتایج در چارچوب پیشینه پژوهش** 111](#_Toc189944775)

[**5-3 بررسی کاربردهای عملی نتایج در دنیای واقعی** 112](#_Toc189944776)

[5-4  **محدودیت‌های پژوهش** 112](#_Toc189944777)

[**5-5 محدودیت‌های ذاتی روش تحقیق یا داده‌های استفاده‌شده** 113](#_Toc189944778)

[**5-6پیشنهادات برای تحقیقات آینده** 113](#_Toc189944779)

[5-6 جمع‌بندی نهایی 114](#_Toc189944780)

[فهرست مراجع 116](#_Toc189944781)

فهرست جدول‌ها

عنوان صفحه

[جدول 2- 1: انواع دیتا بیس ها 32](#_Toc189944786)

[جدول 2- 2: مقایسه مقالات [42], [43], [44] 52](#_Toc189944787)

[جدول 2- 3: ویژگی‌های انتخاب‌شده 54](#_Toc189944788)

[جدول3- 1: مقایسه‌ای دقیق از مدل‌ها 36](#_Toc189944795)

[جدول 3- 2: نمونه ای از داده اولیه 64](#_Toc189944796)

[جدول 3- 3: ویژگی های استخراج شده 64](#_Toc189944797)

[جدول3- 4: پارامترهای رگرسیون لجستیک 69](#_Toc189944798)

[جدول3- 5: پارامترهای مدل جنگل تصادفی 70](#_Toc189944799)

[جدول3- 6: پارامترهای XGB 71](#_Toc189944800)

[جدول 3- 7: پارامترهای تنظیم‌شده 73](#_Toc189944801)

[جدول 4- 1: معیارهای ارزیابی 83](#_Toc189944802)

[جدول 4- 2: تحلیل مقایسه‌ای دو دیتاست 86](#_Toc189944803)

[جدول 4- 3: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست اصلی 91](#_Toc189944804)

[جدول 4- 4: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی 97](#_Toc189944805)

[جدول 4- 5: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی به عنوان داده اموزشی و دادهای اصلی به عنوان داده آزمایشی 103](#_Toc189944806)

فهرست شکل‌ها

عنوان صفحه

[شکل 1- 1: توزيع آسيب‌پذيري‌هاي بحراني برنامه‌هاي وب در سراسر جهان تا سال 2022 [6] 5](#_Toc189944839)

[شکل 1- 2: جايگاه سوم حملات تزريق در 10 مورد اول سال 2021 OWASP Top 10 [7] 5](#_Toc189944840)

[شکل 1- 3: يك سناريو دور زدن احراز هويت[18] 9](#_Toc189944841)

[شکل 2- 1: محبوبيت در بين 500شركت برتر جهان[27] 29](#_Toc189944845)

[شکل 2- 2: تفاوت كوئري پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای 29](#_Toc189944846)

[شکل 2- 3: مدل مفهومی یادگیری ماشین[46] 55](#_Toc189944847)

[شکل 2- 4 : مدل نمونه برای ورودی کاربر مورد نظر حمله[47] 57](#_Toc189944848)

[شکل 2- 5: فلوچارت کد جاوا اسکریپت برای محدود کردن ورودی[29] 58](#_Toc189944849)

[شکل 3- 1: دیاگرام کل کار انجام شده در این پژوهش 61](#_Toc189944851)

[شکل 3- 2: مدل مفهومی روش پیشنهادی 62](#_Toc189944852)

[شکل 3- 3: دیاگرام روش پیشنهادی 62](#_Toc189944853)

[شکل 4- 1: ماتریس همبستگی دیتا مصنوعی 87](#_Toc189944856)

[شکل 4- 2: ماتریس همبستگی دیتا مصنوعی 89](#_Toc189944857)

[شکل 4- 3: ماتریس در هم ریختگی رگرسیون لجستسک 94](#_Toc189944858)

[شکل 4- 4: ماتریس در هم ریختگی درخت تصمیم 94](#_Toc189944859)

[شکل 4- 5: ماتریس در هم ریختگی گرادیان 94](#_Toc189944860)

[شکل 4- 6: ماتریس در هم ریختگی کت نزدیکترین همسایه 95](#_Toc189944861)

[شکل 4- 7: ماتریس در هم ریختگی جنگل تصادفی 95](#_Toc189944862)

[شکل 4- 8: ماتریس در هم ریختگی XGBoost 96](#_Toc189944863)

[شکل 4- 9: ماتریس در هم ریختگی مدل گروهی 96](#_Toc189944864)

[شکل 4- 10: ماتریس در هم ریختگی رگرسیون لجستک 99](#_Toc189944865)

[شکل 4- 11: ماتریس در هم ریختگی درخت تصمیم 100](#_Toc189944866)

[شکل 4- 12: ماتریس در هم ریختگی گرادیان 100](#_Toc189944867)

[شکل 4- 13: ماتریس در هم ریختگی کا نزدیکترین همسایه 101](#_Toc189944868)

[شکل 4- 14: ماتریس در هم ریختگی جنگل تصادفی 101](#_Toc189944869)

[شکل 4- 15: ماتریس در هم ریختگی XGBoost 102](#_Toc189944870)

[شکل 4- 16: ماتریس در هم ریختگی مدل گروهی 102](#_Toc189944871)

[شکل 4- 17: ماتریس در هم ریختگی رگرسیون لجستیک 106](#_Toc189944872)

[شکل 4- 18: ماتریس در هم ریختگی درخت تصمیم 106](#_Toc189944873)

[شکل 4- 19: ماتریس در هم ریختگی گرادیان 106](#_Toc189944874)

[شکل 4- 20: ماتریس در هم ریختگی کا نزدیکترین همسایه 107](#_Toc189944875)

[شکل 4- 21: ماتریس در هم ریختگی جنگل تصادفی 107](#_Toc189944876)

[شکل 4- 22: ماتریس در هم ریختگی XGBoost 108](#_Toc189944877)

[شکل 4- 23: ماتریس در هم ریختگی مدل گروهی 108](#_Toc189944878)

فهرست نمادها

عنوان علامت اختصاري

|  |  |
| --- | --- |
| ضريب ميرايي |  |
|  |  |
|  |  |
|  |  |
|  |  |

کوته‌نوشت‌ها

|  |  |
| --- | --- |
| Additive White Gaussian Noise | AWGN |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

چکيده

تزريق پایگاه‌داده غیررابطه‌ای [[1]](#footnote-2)يك روش حمله است كه عموماً به دلایل ضعف در اعتبارسنجی ورودی‌های برنامه‌ها، سایت‌ها و ابزارها استخراج اطلاعات از پایگاه‌های داده رخ می‌دهد. تحقيقات و كارهاي بسيار در اين حوزه بر روي پایگاه‌داده‌های رابطه‌ای انجام شده است، اما تزريق براي پایگاه‌داده‌های غیررابطه‌ای به دلیل ظهور تازه كمتر موردتوجه قرار گرفته است.

تزريق در پایگاه‌داده به عملياتي اطلاق مي شود كه مهاجم را قادر به قرار دادن عباراتي در پرس وجو می‌کند كه در نهايت با اجراي آن مهاجم ميتواند خواسته هايي را نظير افزايش دسترسي، تغيير در ساختار ديتامدل و دزدي اطلاعات براي خود محقق نمايد.

باتوجه‌به رشد چشمگير استفاده از پایگاه‌داده‌های غیررابطه‌ای به دليل سرعت بالا، توزیع‌پذیری و مزیت‌های مختلف ديگر، توسعه وب‌افزارها با اين دست از پایگاه‌داده افزایش‌یافته و مسئله امنيت اين پایگاه‌داده‌ها بیش‌ازپیش بااهمیت شده است.

هدف از اين پایان‌نامه ارائه راه كاري به‌منظور توليد مجموعه‌داده‌ای استاندارد جهت ارائه به الگوریتم‌های يادگيري ماشين براي كشف منشأ آسیب‌پذیری پایگاه‌داده غیررابطه‌ای با كمترين ميزان خطا و بالاترين كارايي است.

به‌این‌ترتیب در اين پایان‌نامه تلاش شده تا به ابعاد مختلف مسئله تزريق در تحقيق شود. يكي از مهم‌ترین چالش‌ها براي تحقیق‌وتوسعه‌های نوين در اين حوزه فقدان ديتاست قابل‌اطمینان است.

يك توليدكننده مجموعه‌داده به‌منظور استفاده در تحقيقات و پژوهش‌های حوزه يادگيري ماشين و ساير تحقيقات ارائه شده است. اين تلاش شامل دو بخش متادیتاها و ديتاهاي توليد شده است. اين داده‌ها در نهايت در يك مدل يادگيري ماشين، قرار گرفته شده‌اند و به‌عنوان يك دیوار آتش در مقابل حملات پایگاه‌داده غیررابطه‌ای از پایگاه‌داده مربوطه محافظت خواهند نمود.

کلیدواژه‌ها: پایگاه‌داده غیررابطه‌ای، پایگاه‌داده غیررابطه‌ای، تزريق، ديتاست، تشخیص آسیب‌پذیری، مجموعه‌داده [[2]](#footnote-3)، یادگیری ماشینی[[3]](#footnote-4)

# مقدمه

## بيان مسئله

### مسئله و اهداف اصلي تحقيق:

با گسترش روزافزون استفاده از سیستم‌های مبتنی بر وب و پایگاه‌های داده، امنیت اطلاعات و جلوگیری از نفوذ و حملات به این سیستم‌ها از اهمیت ویژه‌ای برخوردار شده است. حملات تزریق داده، ازجمله حملات تزریق پایگاه‌داده رابطه‌ای[[4]](#footnote-5)، به‌عنوان یکی از متداول‌ترین روش‌های سوءاستفاده از آسیب‌پذیری‌های پایگاه‌های داده مطرح هستند[1]. برای مقابله با این نوع حملات، پژوهشگران به تکنیک‌های یادگیری ماشینی روی آورده‌اند. استفاده از این تکنیک‌ها برای شناسایی و جلوگیری از حملات تزریق پایگاه‌داده رابطه‌ای در پژوهش‌های متعددی گزارش شده است و نتایج امیدوارکننده‌ای نیز ارائه داده‌اند.

بااین‌حال، ظهور پایگاه‌های داده پایگاه‌داده غیررابطه‌ای[[5]](#footnote-6) به دلیل قابلیت مقیاس‌پذیری بالا، انعطاف‌پذیری در مدل داده و توانایی مدیریت حجم عظیمی از داده‌ها، باعث گسترش کاربردهای آن در زمینه‌های مختلف شده است[2]. با افزایش استفاده از این نوع پایگاه‌ها، حملات تزریق پایگاه‌داده غیررابطه‌ای به‌عنوان یک چالش امنیتی نوظهور شناخته شده‌اند. برخلاف حملات تزریق پایگاه‌داده رابطه‌ای که بر ساختار جدولی و زبان پرس‌وجوی استاندارد ( پایگاه‌داده رابطه‌ای) متکی هستند، حملات تزریق پایگاه‌داده غیررابطه‌ای از ضعف‌های موجود در ساختارهای غیر جدولی و زبان‌های پرس‌وجوی خاص این نوع پایگاه‌ها سوءاستفاده می‌کنند[3].

علی‌رغم اهمیت فزاینده موضوع، پژوهش‌های کمتری به شناسایی و مقابله با حملات تزریق پایگاه‌داده غیررابطه‌ای پرداخته‌اند. به همین دلیل، توسعه رویکردهای مبتنی بر یادگیری ماشینی برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای به یک نیاز ضروری تبدیل شده است[4]. این تحقیق بر آن است تا با ارائه روش‌هایی نوین، به مقابله با این چالش بپردازد.

یکی از چالش‌های اصلی در اجرای این پژوهش، فقدان مجموعه‌دادگان مناسب و کافی برای آموزش و ارزیابی مدل‌های یادگیری ماشینی است. بسیاری از پژوهش‌های گذشته در زمینه حملات تزریق پایگاه‌داده رابطه‌ای، از مجموعه‌دادگان عمومی و موجود بهره برده‌اند، اما در زمینه حملات پایگاه‌داده غیررابطه‌ای چنین مجموعه‌دادگان استانداردی وجود ندارد. این امر فرایند ارزیابی دقیق و قابل‌اعتماد مدل‌ها را دشوار می‌سازد.

برای غلبه بر این چالش، یکی از روش‌های پیشرفته تولید داده‌های مصنوعی و متنوع، استفاده از مدل‌های زبان بزرگ است. این مدل‌ها، باتکیه‌بر معماری‌های عمیق و مبتنی بر ترانسفورمر، قادر به تولید داده‌هایی با شباهت بالا به داده‌های واقعی هستند. استفاده از مدل‌های زبان بزرگ امکان تولید داده‌های متنی و ساختاریافته را باکیفیت بالا فراهم می‌کند، به‌گونه‌ای که داده‌های تولیدشده قابلیت استفاده به‌عنوان مجموعه‌دادگان آموزشی برای مدل‌های یادگیری ماشینی را دارند. این رویکرد، علاوه بر دقت بالا، انعطاف‌پذیری بیشتری در تطبیق با انواع سناریوها و کاربردها ارائه می‌دهد و می‌تواند جایگزینی مؤثر برای روش‌های سنتی مانند شبکه‌های مولد متخاصم باشد.

## تشريح و بيان موضوع:

برای درک دقیق‌تر موضوع و دست‌یابی به اهداف تحقیق، بررسی و تحلیل دقیق مفاهیم مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای، مجموعه‌دادگان مرتبط، و تکنیک‌های یادگیری ماشینی ضروری است. در این پژوهش، فرضیات و پیش‌فرض‌های زیر به‌عنوان چارچوب اصلی مطالعه و تحلیل در نظر گرفته شده‌اند:

پیش‌فرض‌ها

1. برچسب‌گذاری دستی مجموعه‌داده‌های اولیه:

برای شروع فرایند یادگیری ماشینی، مجموعه‌داده‌های اولیه باید به‌صورت دستی تحلیل و برچسب‌گذاری شوند. این مرحله به‌منظور تضمین کیفیت داده‌های اولیه و فراهم‌کردن مجموعه‌دادگان پایه‌ای برای آموزش مدل‌ها ضروری است.

1. استفاده از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ) برای تولید داده‌های مصنوعی:

به‌منظور گسترش مجموعه‌دادگان و تأمین داده‌های متنوع و واقعی‌تر، از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ) استفاده خواهد شد. این مدل‌ها قادرند داده‌هایی باکیفیت بالا و شباهت زیاد به داده‌های واقعی تولید کنند که می‌تواند به غنی‌سازی مجموعه‌دادگان کمک کند.

1. آزمودن، ارزیابی و دسته‌بندی داده‌ها با یادگیری ماشینی:

برای ارزیابی و دسته‌بندی داده‌های موجود و مصنوعی، از تکنیک‌های مختلف یادگیری ماشینی استفاده خواهد شد. این تکنیک‌ها شامل روش‌های نظارت‌شدن و احتمالی ترکیبی برای شناسایی داده‌های مخرب هستند.

مفاهیم کلیدی پژوهش

بر اساس رویکرد معرفی‌شده، مفاهیم و موضوعات کلیدی زیر در این پژوهش بررسی می‌شوند:

1. تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای:

شناسایی و مقابله با حملات تزریق پایگاه‌داده غیررابطه‌ای، که از نقاط ضعف پایگاه‌داده‌های غیررابطه‌ای استفاده می‌کنند، به‌عنوان هدف اصلی تحقیق بررسی می‌شود. این حملات معمولاً از طریق ارسال درخواست‌های مخرب و دست‌کاری پرس‌وجوها به‌منظور سوءاستفاده از اطلاعات پایگاه‌داده انجام می‌گیرد.

1. بررسی مجموعه‌دادگان موردنیاز

تحلیل و طراحی مجموعه‌دادگان، شامل داده‌های اولیه و داده‌های مصنوعی تولیدشده با مدل‌های زبان بزرگ، یکی از محورهای پژوهش است. این مرحله شامل بررسی کیفیت داده‌ها، شناسایی ویژگی‌های مهم، و طراحی مجموعه‌دادگان قابل‌اعتماد برای مدل‌های یادگیری ماشینی است.

1. بررسی روش‌های تولید داده‌های مصنوعی:

به‌منظور افزایش تنوع و حجم مجموعه‌دادگان، از مدل‌های تولید داده‌های مصنوعی استفاده خواهد شد. در این پژوهش به‌جای استفاده از شبکه‌های مولد متخاصم، از مدل‌های زبان بهره گرفته می‌شود که توانایی تولید داده‌های متنی و ساختاریافته باکیفیت بالا را دارند.

1. طراحی و ارزیابی مدل‌های یادگیری ماشینی:

پژوهش شامل طراحی و ارزیابی یک سیستم یادگیری گروهی خواهد بود که از مدل‌های مختلف مانند Logistic Regression، Random Forest، و XGBoost تشکیل شده است. این مدل‌ها با ترکیب قدرت یادگیری فردی خود، دقت شناسایی حملات را بهبود می‌بخشند.

دستاوردهای مورد انتظار

* ارائه روشی نوین برای شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای:

این پژوهش با معرفی یک رویکرد ترکیبی مبتنی بر یادگیری ماشینی، ابزار قدرتمندی برای شناسایی و مقابله با حملات تزریق پایگاه‌داده غیررابطه‌ای ارائه می‌دهد.

* ایجاد مجموعه‌دادگان جامع و متنوع:

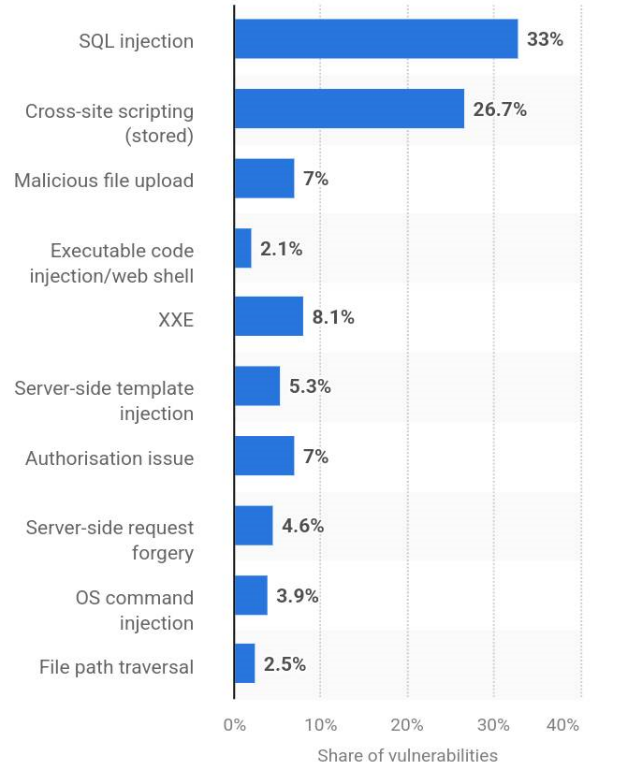
استفاده از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی به غنی‌سازی مجموعه‌دادگان کمک کرده و امکان توسعه مدل‌های دقیق‌تر را فراهم می‌کند.

* بهبود دقت مدل‌های یادگیری ماشینی:

استفاده از مدل‌های ترکیبی و یادگیری گروهی می‌تواند دقت و کارایی سیستم‌های تشخیص را در محیط‌های واقعی به طور قابل‌توجهی افزایش دهد.

### تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای

تزريق پایگاه‌داده رابطه‌ای نوعي حمله سايبري است که برنامه‌هاي مبتني بر داده را هدف قرار می‌دهد[5]. اين اقدام شامل درج عبارات پرس‌وجوي مخرب در فيلدهاي ورودي يا عامل‌هاي يک برنامه وب براي دست‌کاري يا استخراج داده‌هاي حساس از پايگاه‌داده بوده و يكي از آسيب‏پذيري‏هاي اساسي آن‏هاست. در بررسي انجام شده سال 2022 مشخص گرديد كه 33 درصد از آسيب‌پذيري‌هاي حياتي اينترنت، ناشي از اين حملات بوده است. اين امر نشان می‌دهد كه سازمان‌ها نياز فوري به انجام اقدامات پيشگيرانه در مقابله با حملات تزريق پرس‏وجو دارند[6].



شکل 1- 1: توزيع آسيب‌پذيري‌هاي بحراني برنامه‌هاي وب در سراسر جهان تا سال 2022 [6]

در آمار ديگري، از پروژه 10 مورد اول OWASP[[6]](#footnote-7) كه يک استاندارد قابل‌استناد براي توسعه‌دهندگان و كارشناسان امنيت برنامه‌هاي وب بوده و يک توافق جامع در مورد بحراني‌ترين خطرات امنيتي برنامه‌هاي وب است. به جايگاه و اهميت حملات تزريق اشاره شده است[7]



شکل 1- 2: جايگاه سوم حملات تزريق در 10 مورد اول سال 2021 OWASP Top 10 [7]

با اينکه اين آسيب‌پذيري بيش از 20 سال است كه شناخته شده، حملات تزريق هنوز در رتبه سوم آسيب‌پذيري‌هاي وب قرار دارند. در سال 2022، 1162 آسيب‌پذيري با نوع تزريق‌ پرس‏وجو به‌عنوان CVE[[7]](#footnote-8) پذيرفته شده‌اند كه مي‏توان اين برداشت را داشت كه حملات تزريق پرس‏وجو هنوز موجود هستند [8].

در سال 2023، انتظار مي‏رود روند حملات تزريق پرس‏وجو ادامه يابد و فن‏هاي مهاجمان پيچيده‏تر شوند. همانطور که برنامه‏هاي کاربردي وب، پيچيده تر و به هم وابسته‏تر مي شوند، تأثير بالقوه حملات تزريق پرس‏وجو به طور تصاعدي افزايش مي‏يابد[8].

* + بررسي رويكردهاي سنتي مبتني بر امضا

يكي از راه‌حل‌هاي متداول حفظ امنيت در برنامه‏هاي وب استفاده از ديواره‌هاي آتش[[8]](#footnote-9) است. معمولاً اين ديواره‌ها از فن مبتني بر امضا، به‌عنوان هستة اصلي براي شناسايي مخاطرات استفاده مي‌کنند. در اين فن، ديوار آتش هر بسته را در مقابل يک فهرست از "حملات تزريق پرس‏وجو پيش تعريف شده" که به‌عنوان امضاها[[9]](#footnote-10) شناخته مي‌شوند، بررسي مي‌کند. مشکل اين فن آن است که يک حمله‌کننده با دانش خوب از زبان پایگاه‌داده رابطه‌ای مي‌تواند كوئري‌ها را به‌گونه‌اي تغيير دهد که ديوار آتش نتواند آن‌ها را تشخيص دهد، اما همچنان منجر به نتايج خبيثانه مي‌شود [9].

* + بررسي رويكردهاي يادگيري ماشين و يادگيري عميق

نياز به واپايش دقيق‌تر بر اين عامل‌هاي عملياتي با استفاده از روش‌هاي مختلف يادگيري ماشين وجود دارد. در انواع حملات تزريق، شبکه‌هاي عصبي کاربردهاي قابل‌توجهي دارند که شامل بهينه‌سازي، پيش‌بيني، شناسايي، طبقه‌بندي، واپايش، مدل‌سازي و نظارت است. در تحقيقات اخير، بسياري از مسائل حياتي استفاده از يادگيري ماشين و شبکه‌هاي عصبي در عمل مورد بررسي قرار گرفته‌اند. برخي از مشکلات شامل تقسيم داده، جمع‌آوري و مراحل پيش‌پردازش داده‌ها، درنظرگرفتن ورودي‌ها، شبکه‌ها و خروجي‌ها، الگوريتم‌هاي استفاده شده، مدل‌هاي استفاده شده براي آزمون[[10]](#footnote-11) و آموزش[[11]](#footnote-12) و معيارهاي عملکرد تعيين شده در طول اعتبارسنجي و تأييد، هستند[10].

براي نمونه در مدل‌هاي مبتني بر يادگيري عميق، بدون وابستگي به قواعد پايه پايگاه‌داده، از يک مدل پردازش زبان طبيعي[[12]](#footnote-13) و چارچوب يادگيري عميق استفاده مي‌کند. اين روش مي‌تواند دقت را افزايش داده و نرخ هشدارهاي غلط را کاهش دهد درحالي‌که به ماشين اجازه می‌دهد به طور خودکار ويژگي‌هاي مدل زبان حملات تزريق پرس‌وجو را ياد بگيرد، کاهش قابل‌توجهي نيز در دخالت انسان و ارائه دفاع در برابر حملات [[13]](#footnote-14)Zeroday که هرگز رخ نمي‌دهد، فراهم مي‌کند[11].

همه مدل‏هاي موجود به مجموعه‌دادة قابل‌اعتماد و اتكا به‌منظور يادگيري و ارتقا مدل خود نياز دارند، چنان‌که برخي از مدل‏ها، به‌واسطه‌ی تعداد و حجم كم داده آموزشي، غيرقابل‌قبول تلقي مي‏شوند.

### بررسي علل نياز به مجموعه‌داده‏هاي قابل‌اعتماد و قوي

يک مدل يادگيري ماشين عملکرد حداکثري خواهد داشت اگر داده‌هاي آن از نظر كيفي و كمي حد مناسبي داشته باشند. درحالي‌که پژوهشگران بر روي بهبود کيفيت مدل‌ها (مانند جستجوي معماري عصبي و انتخاب خودکار ويژگي) تمرکز داشته‌اند، تلاش‌هاي محدودي براي بهبود کيفيت داده‌ها انجام شده است. يکي از نيازهاي حياتي پيش از استفاده از مجموعه‌داده براي هر كاري، درک مجموعه‌داده آن حوزه است و عدم انجام اين کار ممکن است منجر به تجزيه‌وتحليل نادرست و تصميم‌گيري‌هاي ناقص شود[12].

در خصوص اهميت توسعه مجموعه‏داده، در دنيا نيز كارهاي زيادي انجام شده و يكي از آن‏ها انقلاب معروفي است كه با مجموعه‌داده ImageNet[[14]](#footnote-15) آغاز شد. ازاين‌رو بايد اين ادعا نيز مورد بررسي قرار گرفته مي‌شد. به همين دليل بر روي اين مسئله تحقيقي انجام شد، تا بتوانيم با خيال آسوده‌تري به روند تحقيقاتي خود ادامه دهيم[13].

يادگيري ماشيني كه بر پايه جبر خطي است، داراي تنوع زيادي بوده که مي‌توان در بسياري از حوزه‌ها از آن استفاده کرد، اما براي بهره‏بردن از معماري‌هاي حوزه خاصي مانند واحد پردازش تنسور[[15]](#footnote-16) گوگل محدود است. همچنين، رشد تقاضا براي محاسبات یادگیری ماشین از قانون مور[[16]](#footnote-17) فراتر رفته و نياز است کارشناسان یادگیری ماشین و معماران رايانه براي طراحي سامانه‌هاي محاسباتي موردنياز براي ارائه امکانات یادگیری ماشین به يکديگر کمک کنند. براين‌اساس مي‏توان پيشنهادها و هشدارهايي به معماران رايانه در مورد تحولات يادگيري ماشين ارائه داد[13].

مجموعه‏داده تأثير قابل‌توجهي در افزايش سرعت تحقيقات و توسعه مسائل برنامه‌نويسي دارد. يک مجموعه‌داده براي گسترش تحقيقات در حوزه يادگيري ماشين براي درک بهتر به توليد برنامه كمك می‌کند. دردسترس‌بودن چنين داده‌ها و پايه‌هايي مي‌تواند به توسعه و اعتبارسنجي روش‌هاي جديد کمک کرده و براي مشکلات مختلف درك و توليد نرم‌افزارها استفاده شود[14].

### بررسي اهميت پايگاه‏داده‏هاي غيررابطه‏اي

پايگاه‌هاي اطلاعاتي به سه نوع ساخت‌يافته[[17]](#footnote-18)، نيمه‌ساخت‌يافته[[18]](#footnote-19) و غيرساخت‌يافته[[19]](#footnote-20) طبقه‌بندي مي‌شوند. پايگاه‌داده ساختاريافته نقش‌برجسته‌اي درگذشته داشته است، اما امروزه پايگاه‌داده غيرساخت‌يافته همانند MongoDB[[20]](#footnote-21) رشد قابل‌توجهي را تجربه مي‌کند و در نتيجه، آسيب‌پذيري‌هاي مرتبط با اين پايگاه‌داده نيز در حال افزايش است [15].

طبق يافته‌هاي قبلي مي‏دانيم كه امروزه حملات تزريق از خطرسازترين حملات محسوب مي‌شوند، تحقيقات زيادي در حوزه پايگاه‌داده‌هاي ساخت‌يافته صورت پذيرفته است، كه مي‏توان با تعميم اين كارها به پايگاه‌داده‌هاي ديگر اين مسئله را از ابعاد جديدي مورد بررسي قرارداد.

به همين دليل مجموعه‌داده‌هاي جمع‌آوري‌شده به‌عنوان منبعي ارزشمند براي مطالعه و تحليل آسيب‌پذيري‌هاي تزريق پایگاه‌داده غیررابطه‌ای قابل‌استفاده است كه اطلاعاتي در مورد تهديدات امنيتي بالقوه و کمک به توسعه سازوکارهاي حفاظتي قوي در برابر حملات ارائه می‌دهد. مجموعه‌داده شامل ترکيبي از دستورات پيچيده و ساده بوده که براي يادگيري ماشيني و تجزيه‌وتحليل داده‌ها، به‌ويژه براي علاقه‌مندان به امنيت، سودمند است[15] .

### ارائه نمونه‌ای حمله تزريق پایگاه‌داده غیررابطه‌ای

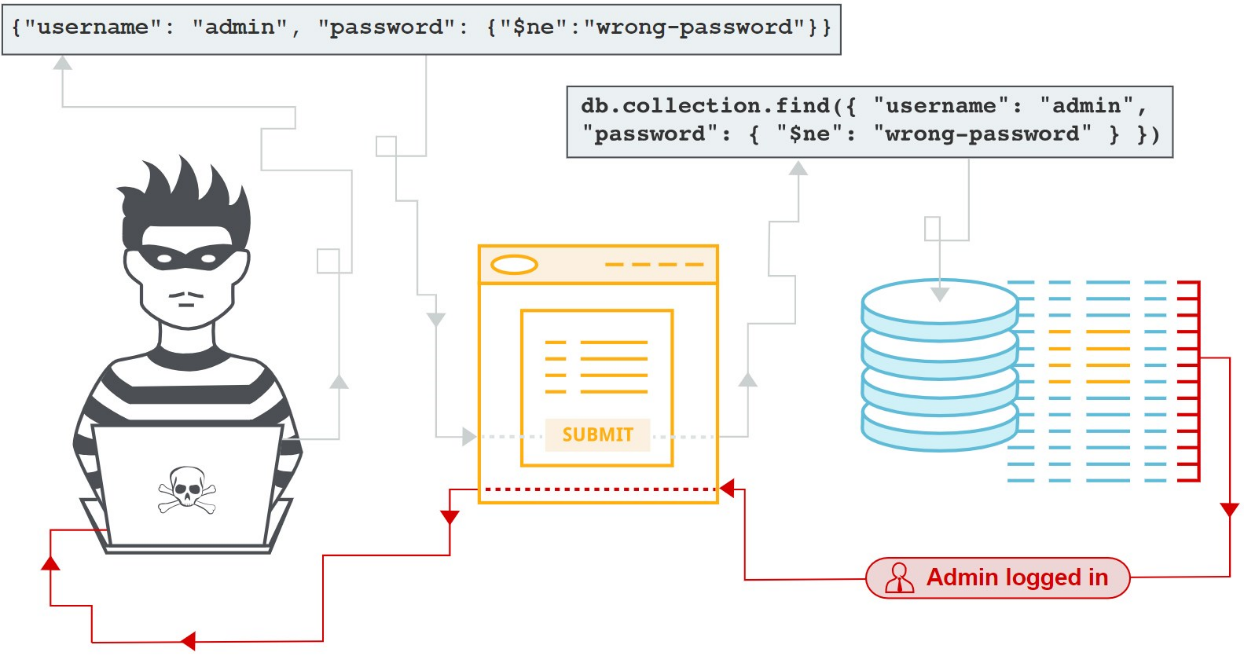
تزريق پایگاه‌داده غیررابطه‌ای آسيب‌پذيري است که در آن يک مهاجم مي‌تواند با استفاده از پرس‏وجوهايي که يک برنامه کاربردي در پایگاه‌داده غیررابطه‌ای دارد تداخل‏هاي ايجاد كرده كه اين تزريق پایگاه‌داده غیررابطه‌ای توانمندي‏هايي نظير زير را به او می‌دهد[16], [17]:

* + دورزدن مکانیسم‌های احراز هويت يا حفاظت
  + استخراج يا ويرايش داده‏ها
  + باعث انکار خدمات
  + اجراکردن کد روی سرور

پایگاه‌داده غیررابطه‌ای داده‏ها را در قالبي غير از جداول سنتي پایگاه‌داده رابطه‌ای ذخيره و بازيابي مي‏کنند. آن‏ها از طيف گسترده‌ای از زبان‌های پرس‌وجو به‌جای يک استاندارد جهاني مانند پایگاه‌داده رابطه‌ای استفاده می‌کنند و محدودیت‌های رابطه‏اي کمتري دارند[16].

دو نوع مختلف تزريق پایگاه‌داده غیررابطه‌ای وجود دارد:

* + تزريق با اشتباه نوشتاري[[21]](#footnote-22): زماني اتفاق مي‌افتد که مي‌توان نوشتار پرس‌وجو پایگاه‌داده غیررابطه‌ای را تغيير داده و محموله[[22]](#footnote-23) خود را تزريق کرد. اين متدولوژي مشابه روشي است که در تزريق پایگاه‌داده رابطه‌ای استفاده می‌شود. بااین‌حال ماهيت حمله به طور قابل‌توجهی متفاوت است، زيرا پايگاه‏هاي داده پایگاه‌داده غیررابطه‌ای از طيف وسيعي از زبان‏هاي پرس‏وجو، انواع نحو پرس‏وجو و ساختارهاي داده متفاوت استفاده مي‏کنند.
  + تزريق عملگر[[23]](#footnote-24) - زماني اتفاق می‌افتد که می‌توان از عملگرهاي پرس‏وجو پایگاه‌داده غیررابطه‌ای براي دست‌کاری پرس‏وجوها استفاده کرد.



شکل 1- 3: يك سناريو دور زدن احراز هويت[18]

در اين پژوهش، نحوه آزمايش آسيب‌پذيري‌هاي پایگاه‌داده غیررابطه‌ای را به‌طورکلی بررسي كرده، سپس بر روي بهره‌برداري از حملات در يكي از انواع پایگاه‌داده غیررابطه‌ای تمرکز خواهد داشت.

به‌عنوان‌مثال مي‏توان اين حملات را در پایگاه‌داده MongoDB مورد بررسي قرارداد تا اين دو نوع حمله بيشتر آشنا شد[19]:

يک وبگاه خريد که محصولات را در دسته‌بندي‌هاي مختلف نمايش می‌دهد را در نظر گرفته. هنگامي که کاربر دسته‌بندي نوشيدني‌هاي گازدار را انتخاب مي‌کند، مرورگر آن‌ها URL زير را درخواست می‌دهد:

|  |
| --- |
| <https://insecure-website.com/product/lookup?category=fizzy> |

اين باعث مي‌شود که برنامه يک پرس‏وجو JSON براي بازيابي محصولات مربوطه از مجموعه محصولات در پایگاه‌داده MongoDB ارسال کند:

|  |
| --- |
| this.category == 'fizzy' |

براي تست اينکه آيا ورودي ممکن است آسیب‌پذیر باشد، یک‌رشته fuzz را در مقدار پارامتر دسته‌بندی ارسال كرده كه یک‌رشته نمونه براي MongoDB به‌صورت زير است:

|  |
| --- |
| ‘”`{  $;Foo}  $Foo \xYZ |

از اين رشته fuzz براي ساخت حمله زير استفاده مي‏شود:

|  |
| --- |
| <https://insecure-website.com/product/lookup?category='%22%60%7b%0d%0a%3b%24Foo%7d%0d%0a%24Foo%20%5cxYZ%00> |

اگر اجرا اين آدرس باعث تغيير در پاسخ اصلي شود، ممکن است نشان‌دهنده اين باشد که ورودي کاربر به‌درستی فيلتر يا تصفيه نشده است.

راه‏هاي مختلف ديگري نيز براي ايجاد بار براي اين دست حملات وجود دارد كه مثال فوق يك نمونه از دست‌کاری در نگارش پرس‏وجو است

در پایگاه‌داده غیررابطه‌ای اغلب از عملگرهاي پرس‌وجو استفاده مي‏شود که راه‏هايي را براي تعيين شرايطي که داده‌ها بايد داشته باشند تا در نتيجه پرس‌وجو قرار گيرند، ارائه می‌دهند[20]. نمونه‌هایی از عملگرهاي پرس‌وجو MongoDB در جدول 1-1 آمده است.

جدول 1- 1: نمونه هايي از عملگرهاي پرس‌وجو MongoDB

|  |  |
| --- | --- |
| با اسنادي مطابقت دارد که عبارت جاوا اسکريپت را برآورده می‌کند. | $where |
| با تمام مقاديري که با يک مقدار مشخص برابر نيستند مطابقت دارد. | $ne |
| با تمام مقادير مشخص شده در يک آرايه مطابقت دارد. | $in |
| اسنادي را انتخاب می‌کند که مقادير با يک عبارت منظم مشخص مطابقت دارند. | $regex |

ممکن است بتوان عملگرهاي پرس‏وجو را براي دست‌کاری آن‏ها در پایگاه‌داده غیررابطه‌ای تزريق كرد. براي انجام اين کار، اپراتورهاي مختلف را به طور سيستماتيک در طيفي از ورودی‌های کاربر ارسال کرده و سپس پاسخ‌ها را براي پیام‌های خطا و يا تغييرات ديگر بررسي بايد كرد.

### بررسي و تشريح مدل

با پیشرفت فناوری هوش مصنوعی و یادگیری عمیق، استفاده از این فناوری‌ها برای شناسایی حملات تزریقی در پایگاه‌های داده به‌ویژه تزریق پایگاه‌داده غیررابطه‌ای به طور چشمگیری افزایش‌یافته است. یکی از چالش‌های اساسی در این حوزه، فراهم‌سازی مجموعه‌داده‌های کافی و مناسب برای آموزش و ارزیابی مدل‌های یادگیری ماشینی است. نبود داده‌های استاندارد و متنوع، استفاده از الگوریتم‌های یادگیری ماشینی را محدود کرده و فرایند طراحی مدل‌های دقیق و مؤثر را دشوار ساخته است.

پژوهش‌های پیشین عمدتاً بر استفاده از روش‌های تقویت داده تمرکز داشته‌اند. این پژوهش‌ها از تکنیک‌هایی نظیر شبکه‌های مولد کانولوشنی عمیق و الگوریتم‌های ژنتیک برای تولید داده‌های مصنوعی و شبیه‌سازی حملات آسیب‌پذیری وب استفاده کرده‌اند. این رویکردها باهدف بهبود دقت مدل‌های طبقه‌بندی، داده‌های متنوعی از حملات تولید کرده و موفقیت‌هایی را در زمینه تشخیص آسیب‌پذیری‌ها به دست آورده‌اند. برای مثال، مدل‌های شبکه‌های مولد متخاصم می‌توانند مجموعه‌داده‌های کوچک را به نمونه‌های بزرگ‌تر و متنوع‌تر گسترش دهند و از این طریق به بهبود عملکرد مدل‌های یادگیری ماشینی کمک کنند.

بااین‌حال، در این پژوهش به‌جای استفاده از شبکه‌های مولد متخاصم، از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی استفاده شده است. این تغییر به دلایل زیر صورت‌گرفته است:

1. توانایی بالای مدل‌های زبان بزرگ در تولید داده‌های متنی واقعی و متنوع:

مدل‌های زبان بزرگ، نظیر GPT، قادر به تولید داده‌هایی با شباهت بالا به داده‌های واقعی هستند که می‌توانند برای شبیه‌سازی سناریوهای مختلف در حملات تزریق پایگاه‌داده غیررابطه‌ای استفاده شوند.

1. انعطاف‌پذیری بیشتر مدل‌های زبان بزرگ در تولید داده‌ها:

برخلاف GAN[[24]](#footnote-25) که به طور خاص برای تولید داده‌های تصویری و ساختاریافته طراحی شده است، مدل‌های زبان بزرگ‌ها می‌توانند داده‌هایی با ساختارهای پیچیده‌تر و کاربردی‌تر در زمینه‌های مختلف از جمله متن، کد و پرس‌وجوهای پایگاه‌داده تولید کنند.

1. سادگی در پیاده‌سازی و تنظیم:

مدل‌های زبان بزرگ نیازی به فرایندهای پیچیده نظیر رقابت بین مولد و تفکیک‌کننده ندارند و تنظیم آن‌ها برای تولید داده‌های دلخواه ساده‌تر است.

رویکرد پیشنهادی برای تولید داده‌ها

در این پژوهش از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی استفاده شده است. فرایند تولید داده‌ها شامل مراحل زیر است[21]:

1. بررسی و تعریف ویژگی‌های موردنیاز:

ابتدا مجموعه ویژگی‌های مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای، نظیر نوع پایگاه‌داده، نوع پرس‌وجو و رفتار مخرب، شناسایی و تعریف می‌شوند.

1. آموزش و تنظیم مدل مدل‌های زبان بزرگ:

مدل مدل‌های زبان بزرگ با استفاده از داده‌های اولیه و برچسب‌گذاری شده، برای تولید داده‌هایی متناسب با سناریوهای حملات تزریق پایگاه‌داده غیررابطه‌ای آموزش داده می‌شود.

1. ارزیابی داده‌های تولیدشده:

داده‌های تولیدشده با استفاده از معیارهای مختلف نظیر شباهت به داده‌های واقعی، تنوع، و پوشش ویژگی‌های مهم ارزیابی می‌شوند.

1. استفاده از داده‌های تولیدشده در مدل‌های یادگیری ماشینی:

داده‌های تولیدشده به‌عنوان بخشی از مجموعه‌داده نهایی برای آموزش مدل‌های یادگیری ماشینی استفاده می‌شوند.

مزایای استفاده از مدل‌های زبان بزرگ در برابر شبکه‌های مولد متخاصم

* تولید داده‌های متنی و ساختاریافته بادقت و کیفیت بالا.
* کاهش پیچیدگی‌های پیاده‌سازی و تنظیم.
* انعطاف‌پذیری بالا در شبیه‌سازی سناریوهای پیچیده.
* پشتیبانی بهتر از تولید داده‌های خاص نظیر پرس‌وجوهای پایگاه‌داده.

دستاوردهای کاربردی

* تولید داده‌های مصنوعی متناسب با حملات پایگاه‌داده غیررابطه‌ای:

مدل‌های زبان بزرگ‌داده‌هایی تولید می‌کنند که شباهت زیادی به داده‌های واقعی دارند و می‌توانند برای شبیه‌سازی حملات استفاده شوند.

* افزایش دقت مدل‌های یادگیری ماشینی:

استفاده از داده‌های مصنوعی تولیدشده با مدل‌های زبان بزرگ، موجب بهبود فرایند آموزش مدل‌های یادگیری ماشینی و افزایش دقت آن‌ها در تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای می‌شود.

* تسهیل فرایند تولید داده‌ها:

این رویکرد زمان و هزینه تولید داده‌های مصنوعی را به طور قابل‌توجهی کاهش داده و نیاز به داده‌های واقعی گسترده را به حداقل می‌رساند[21].

## ضرورت انجام تحقيق:

پژوهش‌های پیشرو در زمینه امنیت سایبری به طور مداوم بر راه‌حل‌های مبتنی بر هوش مصنوعی و یادگیری ماشینی متمرکز بوده‌اند. این موضوع نشان می‌دهد که برای ورود به دنیای هوش مصنوعی و به‌کارگیری ظرفیت‌های آن در حوزه امنیت، دسترسی به مجموعه‌داده‌های جامع، معتبر و مناسب از اهمیت حیاتی برخوردار است. این مجموعه‌داده‌ها نقش کلیدی در آموزش و ارزیابی مدل‌های هوش مصنوعی ایفا می‌کنند و امکان طراحی و پیاده‌سازی روش‌های مؤثرتر را فراهم می‌سازند.

در زمینه حملات تزریق پایگاه‌داده غیررابطه‌ای، ضرورت دسترسی به مجموعه‌داده‌های متناسب دوچندان می‌شود. این حملات به دلیل ساختار غیرمتعارف پایگاه‌داده‌های غیررابطه‌ای و روش‌های خاص تزریق، از چالش‌های نوظهور در حوزه امنیت سایبری به شمار می‌روند؛ بنابراین، شناسایی و پیشگیری از این حملات مستلزم دسترسی به داده‌هایی است که این حملات را به طور دقیق شبیه‌سازی کنند. علاوه بر این، با اشتراک‌گذاری مجموعه‌داده‌های تولیدشده در جامعه امنیت سایبری، می‌توان به پیشرفت‌های چشمگیری در تحقیقات امنیتی دست‌یافت و پایگاه‌داده‌های غیررابطه‌ای را در برابر حملات احتمالی ایمن‌تر کرد.

ضرورت انجام این پژوهش را می‌توان در قالب موارد زیر توضیح داد:

1. توجه به نیازهای بازار کار و صنعت امنیت سایبری:

رشد سریع فناوری‌های مبتنی بر هوش مصنوعی و نیاز روزافزون به متخصصانی که توانایی کار با پایگاه‌داده‌های غیررابطه‌ای و مقابله با حملات سایبری را دارند، انجام این پژوهش را ضروری می‌کند.

1. توسعه سریع نرم‌افزار و اهمیت امنیت در چرخه توسعه:

با افزایش تقاضا برای توسعه سریع نرم‌افزارهای مبتنی بر پایگاه‌داده‌های غیررابطه‌ای، حفظ امنیت این سیستم‌ها در برابر حملات سایبری نظیر تزریق پایگاه‌داده غیررابطه‌ای به یکی از الزامات مهم در چرخه توسعه نرم‌افزار تبدیل شده است.

1. تقاضای جهانی و داخلی برای اینترنت و خدمات آنلاین:

افزایش وابستگی به خدمات آنلاین که به پایگاه‌داده‌های غیررابطه‌ای متکی هستند، نیاز به ایمن‌سازی این سیستم‌ها را به‌عنوان یک ضرورت جهانی مطرح می‌کند.

1. اهمیت هوش مصنوعی در امنیت سایبری:

هوش مصنوعی نقش بی‌بدیلی در شناسایی و پیشگیری از حملات سایبری دارد. مدل‌های هوش مصنوعی پیشرفته می‌توانند با تحلیل الگوهای رفتاری در داده‌ها، حملات احتمالی را پیش‌بینی کرده و از وقوع آن‌ها جلوگیری کنند.

1. اهمیت و پیچیدگی حملات تزریق:

حملات تزریق داده، به‌ویژه تزریق پایگاه‌داده غیررابطه‌ای، به دلیل آسیب‌پذیری‌های خاص پایگاه‌داده‌های غیر جدولی، پیچیدگی بیشتری نسبت به حملات سنتی پایگاه‌داده رابطه‌ای دارند. بررسی و مقابله با این حملات نیازمند پژوهش‌های عمیق و مجموعه‌داده‌های متناسب است.

تأثیر بر کاربران و جامعه سایبری

همه ما خواستار حضور امن در فضای سایبری هستیم. باتوجه‌به گسترش فعالیت‌های کاربران در این فضا، امنیت اطلاعات و سیستم‌های مبتنی بر پایگاه‌داده‌های غیررابطه‌ای به یکی از نیازهای اصلی تبدیل شده است. این پژوهش می‌تواند از جنبه‌های زیر به بهبود امنیت سایبری کمک کند:

1. کاهش خطرات ناشی از نقض داده‌ها:

شناسایی دقیق حملات تزریق پایگاه‌داده غیررابطه‌ای به کاهش احتمال نقض داده‌ها و حملات سایبری در پایگاه‌داده‌های مدرن کمک می‌کند.

1. ایجاد اعتماد در استفاده از پایگاه‌داده‌های غیررابطه‌ای

ارائه راه‌حل‌های امنیتی قابل‌اعتماد، باعث افزایش اعتماد کاربران و سازمان‌ها به استفاده از این پایگاه‌داده‌ها خواهد شد.

1. تقویت جامعه امنیت سایبری از طریق اشتراک‌گذاری مجموعه‌داده‌ها:

انتشار مجموعه‌داده‌های مصنوعی تولیدشده با استفاده از مدل‌های زبان بزرگ، امکان انجام تحقیقات گسترده‌تر و توسعه روش‌های جدید برای مقابله با حملات سایبری را فراهم می‌کند.

ضرورت انجام این پژوهش نه‌تنها در حوزه نظری بلکه در کاربردهای عملی آن نیز مشهود است. این تحقیق با تمرکز بر استفاده از مدل‌های پیشرفته یادگیری ماشینی و تولید داده‌های مصنوعی متناسب، گامی مؤثر در جهت افزایش امنیت سایبری و مقابله با تهدیدات نوظهور در پایگاه‌داده‌های غیررابطه‌ای خواهد بود.

### بر مبناي تمايل بازار كار:

تقاضا براي مهارت‌هاي مديريت پايگاه‌داده فراتر از شرکت‌هاي وب يا نرم‌افزار و به صنايعي مانند خرده‌فروشي، بيمارستان‌ها و حتي دولتي گسترش‌يافته است. اين صنايع به دنبال افرادي بامهارت در مديريت و تجزيه‌وتحليل مجموعه‌داده‌هاي بزرگ[[25]](#footnote-26) هستند و وقتي صحبت از مهارت‌هاي موردنظر مي‌شود، جاي تعجب نيست که دانش پایگاه‌داده غیررابطه‌ای و Hadoop[[26]](#footnote-27) بسيار موردنياز است [22].

تمايل بازار کار طبيعتاً برگرفته از حركت فنّاوري و نيازسنجي‌هاي مختلف آن بوده و استفاده از پایگاه‌داده غیررابطه‌ای‌ها در بازارهاي كار مبتني بر اينترنت اشيا[[27]](#footnote-28) اجتناب‌ناپذير است.

رشد تعاملات متقابل بين اينترنت اشيا و فناوري‌هاي اجتماعي[[28]](#footnote-29) به پيشرفت‌ در سامانه‌هاي اجتماعي سايبر فيزيکي[[29]](#footnote-30) منجر شده است. يکي از چالش‌هاي اصلي اين ادغام فناوري مربوط به امنيت اطلاعات[[30]](#footnote-31) و حريم خصوصي[[31]](#footnote-32) است. شناسايي تعاملات مربوطه و چالش‌هاي اصلي که به طور عمده بر روي امنيت سايبري و حريم خصوصي تمرکز دارد. مشارکت‌هاي خاص شامل:

1. تحول ديجيتال و تأثير آن بر ادبيات تجاري
2. شناسايي فرصت‌هاي مربوط به ادغام فناوري‌هاي اجتماعي و اينترنت اشيا
3. چالش‌هاي مربوط به همگرايي فناوري با تمرکز بر چالش‌هاي امنيت سايبري [21]

### بر مبناي نياز به توسعه سريع:

مدل‌هاي مختلف توسعه نرم‌افزار با رويكردهاي چابك[[32]](#footnote-33) برنامه‌نويسان زيادي را بر آن داشته كه بجاي تحليل ديتا مدل‌هاي پيچيده به سمت استفاده از پايگاه‌داده‌هاي غيررابطه‌اي بروند تا به‌راحتي بتوانند داده‌ها را با ديد انتزاعي‌تري در قالب‌هاي JSON[[33]](#footnote-34), X یادگیری ماشین[[34]](#footnote-35) و غيره ذخيره و استفاده نمايند.

از طرفي الگوهاي طراحي نظير CQRS[[35]](#footnote-36) پديد آمدند كه به استفاده از چند نوع پايگاه‌داده توصيه مي‌کنند، كه در اين صورت مي‌توان از قابليت‌هاي متنوع آن‏ها بنا به نياز در كمترين زمان، توسعه وب‌افزارهاي مختلف را ادامه داد.

### بر مبناي تقاضاي جهاني و داخلي:

تا ژانويه 2022، كاربران جهاني اينترنت به ۷.۹۳ ميليارد نفر رسيده است و اتحاديه بين‌المللي مخابرات پيش‌بيني مي‌کند که تا پايان سال ۲۰۲۴، حدود ۶۷.۹ درصد جمعيت جهان به اينترنت دسترسي پيدا کنند[22]. در ژانويه 2022، تعداد کاربران اينترنت در ايران 78,086,663 نفر بود که معادل 91.0٪ از جمعيت کشور است[22]. همگام با اين رشد فوق العاده از کاربران اينترنت، رشد مشابهي در ارائه خدمات از طريق اينترنت وجود داشته است. تاکنون، خدمات آنلاين متعددي از جمله تجارت الکترونيک، پشتيباني تجارت‏ها با تجارت‏هاي نوين و مخازن بزرگ داده معرفي شده‏اند همچنين بسياري از خدمات نوآورانه ديگر هنوز در آينده نزديک ارائه خواهند شد. با افزايش تعداد سرويس‌هاي آنلاين، تهديدات امنيتي نيز افزايش مي‌يابد، که در حال حاضر هم براي کاربران اينترنت و هم براي ارائه‌دهندگان خدمات آنلاين به يک نگراني بزرگ تبديل شده است. با در دسترس داشتن منابع سازمان به صورت آنلاين و اتخاذ تدابير امنيتي سطح پايين، اشخاص غيرمجاز مي‏توانند به داده‏هاي محرمانه و حساس يک سازمان دسترسي داشته باشند. امروزه پايگاه‌داده نرم‏افزار‏هاي وب به هدف اصلي اين نوع دسترسي‏هاي غيرمجاز تبديل شده‏اند[23].

### بر مبناي اهميت هوش مصنوعي

هوش مصنوعي[[36]](#footnote-37) مي‌تواند براي جامعه مفيد باشد، اما همچنين مملو از خطرات است. پذيرش اجتماعي هوش مصنوعي به اعتماد ذي‌نفعان به هوش مصنوعي بستگي دارد، بااين‌حال ادبيات مربوط به اعتماد به هوش مصنوعي ناقص است، و اطلاعات کمي در مورد آسيب‌پذيري‌هايي که ذي‌نفعان مختلف با آن مواجه هستند شناخته شده كه انتقال آن براي عموم آسان نيست [24].

### بر مبناي اهميت حملات تزريق

تزريق پرس‌وجو به دسته‌اي از انواع حملات اشاره دارد که شامل درج عامل‌هاي پرس‌وجوي مخرب ساخته شده در عمليات پرس‌وجو پايگاه‌داده و يا فراخواني‌هاي [[37]](#footnote-38)API است. اگرچه اين مسائل امنيتي به طور گسترده در پايگاه‌هاي اطلاعاتي رابطه‌اي موردمطالعه قرار گرفته‌اند، امکان و بروز تزريق در ذخيره‌سازهاي داده پایگاه‌داده غیررابطه‌ای که بر اساس مدل‌هاي داده متفاوت ساخته شده‌اند، کمتر موردتوجه قرار گرفته است[25].

حملات تزريق پایگاه‌داده غیررابطه‌ای يك روش حمله است كه موجب استخراج اطلاعات از پايگاه‌داده‏ شده و عموماً به دلايل ضعف در اعتبارسنجي ورودي‌ برنامه‌ها، وبگاه‌ها و ابزارها رخ می‌دهد. تحقيقات و كارهاي بسيار در اين حوزه بر روي پايگاه‌داده‌هاي رابطه‌اي انجام شده است. تزريق براي پايگاه‌داده‌هاي غيررابطه‌اي به دليل نوظهور بودن آن‏ها، كمتر موردتوجه قرار گرفته است. همچنين توسعه و تعميم روش‌ها مختلف اين حوزه، خصوصاً با استفاده از مدل‌هاي يادگيري ماشين، يادگيري عميق و شبکه‌هاي عصبي نياز به حجم مناسبي از مجموعه‏داده تأييد شده، معتبر و قابل‌استناد، بسيار الزامي است.

باتوجه‌به رشد چشمگير استفاده از پايگاه‌داده‌هاي غيررابطه‌اي به دليل سرعت بالا، توزيع‌پذيري و وب‌افزارها مختلف ديگر، توسعه وب‌افزارها با اين دست از پايگاه‌داده افزايش‌يافته و مسئله امنيت اين پايگاه‌داده‌ها بيش‌ازپيش بااهميت شده است. از طرفي اهميت اين تحقيقات براي كمك به‌ فايروال‏ها به‌عنوان اولين لايه دفاعي که به آن حمله مي‌شود، بر کسي پوشيده نيست. برنامه‏هاي کاربردي فايروال وب[[38]](#footnote-39) نقشي ضروري در دفاع برابر حملات وب مخرب مانند تزريق كوئري[[39]](#footnote-40) ايفا مي‌کنند. با توسعه محاسبات ابري[[40]](#footnote-41)، برنامه‏هاي کاربردي فايروال وب به‌عنوان يکي از انواع امنيت، به‌عنوان سرويس تحت وب پيشنهاد شده است. براي تسهيل استقرار، پيکربندي و به‌روزرساني برنامه‏هاي کاربردي ديواره آتش وب‏، اما باوجود محبوبيت فوق‌العاده آن، آسيب‌پذيري‌ برنامه‏هاي کاربردي فايروال وب به‌عنوان سرويس، هنوز تا حد زيادي ناشناخته بوده که باتوجه‌به استفاده گسترده آن، بسيار نگران‌کننده است[25].

## سابقه تحقيقات و مطالعات انجام‌گرفته

با بررسي در مطالعات انجام شده متناسب با اين پژوهش به دسته‌بندي از مقالات اصلي خواهيم رسيد:

* تشخيص حملات تزريق پایگاه‌داده غیررابطه‌ای

در پايان‌نامه آقاي رومياني در سال 1402 با استفاده از مدل‏هاي يادگيري ماشين در خصوص نحوه پيشگيري از حملات تزريق راه‌حل‌هايي ارائه شده است. در جلسات دفاع نيز نسبت به کم‌بودن حجم مجموعه‌داده‏ و دشواري زيادي كه ايشان براي به‌دست‌آوردن مجموعه‏داده داشتند، اشاره شد [26].

آقاي داس[[41]](#footnote-42) و همكاران در سال 2019 نيز در خصوص آسيب‌پذيري از تزريق پرس‌وجو كار كردند، آن‏ها بر اين باور هستند كه هر زمان که برنامه وب دستورات پایگاه‌داده رابطه‌ای پويا[[42]](#footnote-43) را اجرا مي‌کند ممکن است تحت حمله تزريق پرس‏وجو قرار گيرد. براي ارزيابي شيوه‌هاي موجود تشخيص آن، ما دو سناريو امنيتي مختلف را براي احراز هويت برنامه‌هاي وب در نظر مي‌گيريم که پرس‌وجوي پایگاه‌داده رابطه‌ای پويا را با داده‌هاي ورودي کاربر ايجاد مي‌کند. براين‌اساس، دو مجموعه‌داده مختلف را با درنظرگرفتن تمام آسيب‌پذيري‌هاي احتمالي در فواصل زماني اجرا توليد نموده. رويکرد پيشنهادي مبتني بر فاصله ويرايش براي طبقه‌بندي يک پرس‌وجوي پایگاه‌داده رابطه‌ای پويا به‌عنوان عادي يا مخرب با استفاده از نمايه وب آماده‌شده با پرس‌وجوهاي پایگاه‌داده رابطه‌ای پويا در طول مرحله آموزش ارائه مي‌شود. مجموعه‌داده را با استفاده از رويکرد پيشنهادي و برخي از رويکردهاي طبقه‌بندي نظارت شده معروف، ارزيابي نموده كه در نتيجه روش آن‏ها در تشخيص حمله تزريق پرس‏وجو تحت هر دو سناريو امنيت احراز هويت مؤثرتر است [27].

در حوزه حملات تزريق پایگاه‌داده غیررابطه‌ای نيز مي‏توان به تلاش آقاي لاندايت و همكاران در سال 2024 اشاره داشت، در اين مقاله، نتايج يک بررسي عميق از خطرات مربوط به تزريق در پايگاه‌داده گراف[[43]](#footnote-44) Neo4j و اکوسيستم گسترده‌تر آن را ارائه شده كه بر اساس دو معيار تقسيم شده است. اول، بررسي اجراي توزيع شده پرس‌وجوهاي عاملي[[44]](#footnote-45)، از اتصال‌دهنده‌هاي زبان ويژه كاربران[[45]](#footnote-46)، تا ارتباطات (پروتکل Bolt) و اجرا در Neo4j (در طرح‌هاي پرس‌وجو)، و دوم شناسايي مشکلات تزريق باقيمانده در مواردي که پرس‌وجوهاي ثابت و عاملي شده کافي نيستند. اين مطالعه شامل بررسي جريان داده‌هاي کد محور با پايه کد Neo4j است و با مجموعه آزمايشي موارد آزمون تزريق تکميل مي‌شود.

* تأثیر توليد مجموعه‌داده‏ها

آقاي يانگ[[46]](#footnote-47) در سال 2019 به يك انقلاب در حوزه رايانه اشاره دارند كه با مجموعه‏داده ImageNet رخ داد، براي اين تعريف جهت‏گيري‏هايي نيز تعريف كرده‏اند كه انقلاب فعلي یادگیری ماشین به دو نوع مقياس نياز دارد:

1. در مجموعه‌داده‌هاي موجود
2. منابع محاسباتي مورداستفاده براي تجزيه‌وتحليل آن‏ها.

مجموعه‌داده‌هاي بزرگ حاوي مواد خام براي درک دنياي اطراف ما هستند. تا کنون، محاسبات یادگیری ماشین در مقياس بزرگ در ديتاسنترهاي بزرگ حاوي انبوهي از GPUها که در اصل براي تسريع گرافيک طراحي شده بودند، انجام شده است. ما در حال گذار به عصري هستيم که در آن ديتاسنترها پر از رايانه‌هايي مي‌شوند که صرفاً براي محاسبات یادگیری ماشین طراحي شده‌اند. آن‏ها نه‌تنها از توليدکنندگان GPU، بلکه از استارتاپ‌ها، از خطوط توليد جديد عرضه‌کنندگان ريزپردازنده‌هاي سنتي، و از شرکت‌هاي اينترنتي معتبرتر که قبلاً پردازنده‌هاي خود را ساخته بودند، ارائه مي‌شوند.

* + توليد مجموعه‌داده براي پایگاه‌داده رابطه‌ای

آقاي امين اقباليان و همكاران در سال 1401 در پژوهشي به نام توليد مجموعه‌داده استاندارد جهت تشخيص آسيب‌پذيري‌هايي منشأ پایگاه‌داده رابطه‌ای I براي استفاده در راهكارهاي مبتني بر يادگيري ماشيني مجموعه‌دادة براي اين منظور توليد كردند. اين مقاله براي نيل به اين منظور رفتار يك هكر واقعي و اطلاعاتي كه يك هكر جمع‏آوري می‌کند ملاك عمل براي توليد اين مجموعه‏داده قرار گرفته است. پس از ارزيابي، ميزان دقت كلي راهكار پيشنهادي توليد اين مجموعه‏داده براي تشخيص آسيب‏پذيري برنامه تحت آزمون %٩٨ درصد است.

* توليد مجموعه‌داده براي حملات تزريق پایگاه‌داده غیررابطه‌ای

آقاي رامشوار[[47]](#footnote-48) و ساير محققين در سال 2023، اذعان داشتند كه در خصوص پایگاه‌داده رابطه‌ای i كارهاي بسياري انجام شده است. اما آنچه در اين پژوهش به طور ويژه موردبحث است حملات تزريق پایگاه‌داده غیررابطه‌ای است كه مقاله فوق به توليد 400 نمونه داده آن هم صرفاً براي MongoDB پرداخته است. ازاين‌رو يک مجموعه‌داده جامع شامل 400 دستور تزريق پایگاه‌داده غیررابطه‌ای جمع‌آوري‌شده است. اين دستورات به دودسته تقسيم مي‌شوند: 221 فرمان مخرب و 179 فرمان خوش‌خيم. مجموعه‌داده به‌دقت با ترکيب دستورات نوشته شده به‌صورت دستي و دستورات به‌دست‌آمده از طريق وب از منابع معتبر مانند GitHub و ساير وبگاه‌هاي مشخص شده که در بخش منابع اين مقاله توضيح داده شده است، تنظيم شده است. مجموعه‌داده‌هاي جمع‌آوري‌شده به‌عنوان منبعي ارزشمند براي مطالعه و تحليل آسيب‌پذيري‌هاي تزريق پایگاه‌داده غیررابطه‌ای، ارائه بينش‌هايي در مورد تهديدات امنيتي بالقوه و کمک به توسعه سازوکارهاي حفاظتي قوي در برابر چندين حمله عمل مي‌کند. مجموعه‌داده شامل ترکيبي از دستورات پيچيده و ساده است که هنوز براي يادگيري ماشيني و تجزيه‌وتحليل داده‌ها، به‌ويژه براي علاقه‌مندان به امنيت، مناسب است [10].

* توليد مجموعه‌داده مبتني بر GAN

در تحقيق آقاي ون‏ژو[[48]](#footnote-49) و محققين ديگر، مسئله کمبود داده‌هاي تجاري در سامانه واپايش صنعتي موردمطالعه قرار گرفت و الگوريتم توليد نمونه حمله ارائه شده، ابتدا، نتايج گروه‌بندي همبستگي به‌وسيله توزيع وزن و عضويت به دست مي‌آيد و سپس نتايج گروه‌بندي ارتباط قوي براي حمله به‌دست‌آوردن نمونه‌هاي حمله مورد حمله قرار مي‌گيرد. در نهايت، از GAN براي گسترش نمونه استفاده مي‌شود. اين مقاله از مجموعه‌داده‌هاي باز و يک مجموعه‌داده از نفتکش ساختمان نمونه‌هاي حمله، درجه تطابق و سازگاري را توليد مي‌کند [13].

در اين مقاله، نمونه‌هاي حمله توسط مجموعه‌داده‌هاي [[49]](#footnote-50)BATADAL و مجموعه‌داده‌هاي تجاري يک انبار نفت توليد مي‌شوند و داده‌ها از طريق الگوريتم 100 برابر گسترش مي‌يابند [13].

همچنين در تحقيق آقاي لودانگ‏ژو[[50]](#footnote-51) و همكاران، با دانش به اينكه، انتخاب مجموعه‏داده نمونه، عامل تعيين‌کننده‌اي است و پاسخ به سؤال: آيا الگوريتم‌هاي هوش مصنوعي مي‌توانند در آن به نتايج خوبي دست يابند؟ به سراغ توليد مجموعه‏داده تزريق با مدل مبتني بر GAN اقدام شد [12].

اما مجموعه‌داده‌ها با برچسب‌هاي دسته‌بندي خاص به‌سختي به دست مي‌آيند، در اين مقاله بر تقويت داده‌ها براي يادگيري نمايش ويژگي‌هاي مشابه از داده‌هاي اصلي براي بهبود دقت مدل‌هاي طبقه‌بندي تمرکز كرده و شبکه‌هاي متخاصم مولد کانولوشني عميق همراه با الگوريتم‌هاي ژن‌شناختي در زمينه حملات آسيب‌پذيري وب باهدف حل مشکل تعداد ناکافي نمونه‌هاي تزريق پرس‏وجو اعمال شده است. همچنين انتظار مي‌رود اين روش براي توليد نمونه براي انواع ديگر حملات آسيب‌پذيري نيز اعمال شود .[12]

شبکه متخاصم مولد[[51]](#footnote-52) متشکل از مولدها[[52]](#footnote-53) و تمايز کنندگاني[[53]](#footnote-54) است که عليه يکديگر بازي مي‌کنند. داده‌هاي ورودي با نمونه‌گيري تصادفي در فضاي پنهان[[54]](#footnote-55) به‌دست‌آمده و به شبکه مولد منتقل مي‌شود. شبکه مولد، يک رويکرد مبتني بر سنتز[[55]](#footnote-56) شبکه، نمونه‌هاي متنوع‌تري را در مقايسه با تكنيك‌هاي سنتي افزايش داده توليد مي‌کند، اگرچه فرايند آن پيچيده‌تر است [12].

## هدف پژوهش

هدف اصلی این پژوهش، ارائه رویکردی جامع و نوین برای مقابله با تهدیدات امنیتی مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای است. این هدف با طراحی و توسعه یک سیستم مبتنی بر یادگیری ماشینی، باقابلیت تشخیص دقیق و کارآمد بارهای مخرب [[56]](#footnote-57)و تمایز آن‌ها از بارهای قانونی[[57]](#footnote-58)، دنبال می‌شود.

جزئیات اهداف تحقیق

1. ایجاد و استفاده از یک مجموعه‌داده جامع:

برای اطمینان از دقت و کارایی مدل‌های یادگیری ماشینی، پژوهش بر تولید و استفاده از یک مجموعه‌داده جامع و متنوع تمرکز دارد. این مجموعه‌داده با بهره‌گیری از مدل‌های زبان بزرگ تولید شده است که امکان شبیه‌سازی دقیق سناریوهای مختلف حملات تزریق پایگاه‌داده غیررابطه‌ای را فراهم می‌کند.

1. انتخاب و استخراج ویژگی‌های مؤثر:

شناسایی و انتخاب ویژگی‌های کلیدی از داده‌ها که به شناسایی الگوهای مخرب کمک می‌کنند، یکی از گام‌های مهم در این پژوهش است. این ویژگی‌ها شامل مشخصه‌های ساختاری، رفتاری و محتوایی بارهای داده هستند.

1. طراحی و بهینه‌سازی مدل‌های یادگیری ماشینی:

مدل‌های پیشنهادی، شامل Logistic Regression، Random Forest، و XGBoost، به‌صورت مجزا و در یک سیستم یادگیری گروهی استفاده می‌شوند. بهینه‌سازی این مدل‌ها از طریق تنظیم پارامترها و روش‌های معتبر ارزیابی، بهبود دقت و کاهش نرخ خطا را تضمین می‌کند.

1. بررسی رویکرد پیشنهادی از منظر امنیت و کارایی:

رویکرد ارائه‌شده از دو منظر زیر ارزیابی می‌شود:

* + امنیت: توانایی مدل در شناسایی دقیق حملات تزریق پایگاه‌داده غیررابطه‌ای و جلوگیری از نفوذ مخرب.
  + کارایی: عملکرد مدل از نظر زمان پردازش، منابع موردنیاز، و قابلیت مقیاس‌پذیری در سناریوهای واقعی.

1. مقایسه با روش‌های موجود:

رویکرد پیشنهادی با روش‌های موجود در زمینه شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای مقایسه می‌شود تا مزایا و محدودیت‌های آن به‌صورت دقیق ارزیابی شوند. این مقایسه بر اساس معیارهایی نظیر صحت، نرخ مثبت کاذب (False Positive Rate)، و زمان پردازش انجام می‌شود.

پژوهش حاضر بر استفاده از مدل‌های یادگیری ماشینی متکی است که با استفاده از مجموعه‌داده‌های تولیدشده توسط مدل‌های زبان بزرگ آموزش داده می‌شوند. این مجموعه‌داده‌ها شامل داده‌های مصنوعی با ویژگی‌هایی شبیه به داده‌های واقعی هستند و نقش کلیدی در بهبود عملکرد مدل‌ها ایفا می‌کنند. تنظیم دقیق پارامترهای مدل و استفاده از تکنیک‌های پیشرفته نظیر یادگیری گروهی، منجر به بهبود قابل‌توجه در دقت تشخیص حملات و کاهش نرخ هشدارهای کاذب می‌شود.

دستاوردهای مورد انتظار پژوهش:

* دقت بالا در تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای:

با استفاده از مجموعه‌داده‌های جامع و مدل‌های بهینه‌سازی‌شده، انتظار می‌رود که مدل پیشنهادی دقت بالایی در شناسایی حملات داشته باشد.

* کاهش نرخ هشدارهای کاذب:

بهبود الگوریتم‌ها و تنظیم دقیق پارامترها، منجر به کاهش هشدارهای کاذب و افزایش اعتماد به نتایج مدل خواهد شد.

* افزایش امنیت پایگاه‌داده‌های غیررابطه‌ای:

رویکرد پیشنهادی می‌تواند به بهبود امنیت پایگاه‌داده‌های غیررابطه‌ای و کاهش خطرات ناشی از حملات سایبری کمک کند.

* مقایسه دقیق با روش‌های موجود:

تحلیل و مقایسه نتایج با روش‌های موجود، جایگاه علمی و کاربردی رویکرد پیشنهادی را مشخص کرده و مسیر تحقیقات آینده را روشن‌تر می‌کند.

## نوآوري پژوهش

این پژوهش باهدف ارائه رویکردی جدید برای شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای، چندین جنبه نوآورانه را دنبال می‌کند که از لحاظ علمی و کاربردی در حوزه امنیت سایبری و پایگاه‌داده‌های غیررابطه‌ای برجسته هستند. نوآوری‌های این پژوهش را می‌توان در قالب موارد زیر دسته‌بندی کرد:

استفاده از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی

* برخلاف پژوهش‌های پیشین که عمدتاً از روش‌هایی نظیر شبکه‌های مولد متخاصم یا الگوریتم‌های کلاسیک برای تولید داده‌های مصنوعی استفاده کرده‌اند، این پژوهش از مدل‌های زبان بزرگ برای تولید داده‌های مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای بهره می‌گیرد.
* مدل‌های زبان بزرگ قادرند داده‌هایی با شباهت بالا به داده‌های واقعی تولید کنند که کیفیت و دقت مجموعه‌داده‌ها را افزایش می‌دهند. این روش نوآورانه امکان تولید داده‌های متنی پیچیده و غنی‌تری را فراهم می‌کند.

طراحی یک سیستم یادگیری گروهی برای شناسایی حملات

* این پژوهش از ترکیب چند مدل یادگیری ماشینی شامل Logistic Regression، Random Forest و XGBoost در قالب یک سیستم یادگیری گروهی استفاده می‌کند.
* این ترکیب، توانایی مدل‌ها در تشخیص حملات را تقویت کرده و به بهبود دقت، کاهش نرخ هشدارهای کاذب و افزایش کارایی مدل‌ها کمک می‌کند.

تمرکز بر حملات تزریق پایگاه‌داده غیررابطه‌ای به‌عنوان یک چالش نوظهور

* پژوهش‌های پیشین بیشتر بر حملات تزریق پایگاه‌داده رابطه‌ای تمرکز داشته‌اند، درحالی‌که این تحقیق به طور خاص به حملات تزریق پایگاه‌داده غیررابطه‌ای می‌پردازد که به دلیل تفاوت ساختاری پایگاه‌داده‌های غیررابطه‌ای، به روش‌های امنیتی متفاوت و پیچیده‌تری نیاز دارند.
* بررسی عمیق این نوع حملات و ارائه رویکردی خاص برای شناسایی آن‌ها، خلأ موجود در پژوهش‌های امنیت سایبری را پر می‌کند.

طراحی و ایجاد یک مجموعه‌داده جامع و نوآورانه

* یکی از مهم‌ترین نوآوری‌های این پژوهش، ایجاد یک مجموعه‌داده مصنوعی و متنوع برای حملات تزریق پایگاه‌داده غیررابطه‌ای است. این مجموعه‌داده با استفاده از مدل‌های زبان بزرگ تولید شده و به‌عنوان منبعی قابل‌اعتماد برای آموزش و ارزیابی مدل‌های یادگیری ماشینی در این حوزه عمل می‌کند.
* با اشتراک‌گذاری این مجموعه‌داده در جامعه تحقیقاتی، امکان توسعه روش‌های جدید و انجام تحقیقات بیشتر در این زمینه فراهم می‌شود.

رویکرد جامع برای ارزیابی امنیت و کارایی

* در این پژوهش، رویکرد پیشنهادی از دو جنبه امنیت و کارایی مورد ارزیابی قرار می‌گیرد:
  + امنیت: توانایی مدل در تشخیص حملات مخرب و جلوگیری از نفوذ.
  + کارایی: بررسی زمان پردازش، منابع موردنیاز، و قابلیت مقیاس‌پذیری رویکرد در سناریوهای واقعی.
* این ارزیابی جامع کمک می‌کند تا رویکرد پیشنهادی به‌عنوان یک راه‌حل عملی و کاربردی در محیط‌های واقعی مورداستفاده قرار گیرد.

مقایسه و تحلیل جامع با روش‌های موجود

* این تحقیق رویکرد پیشنهادی را با روش‌های موجود از جنبه‌های مختلف نظیر دقت، نرخ مثبت کاذب و زمان پردازش مقایسه می‌کند.
* تحلیل نتایج مقایسه‌ای به شناسایی مزایا و محدودیت‌های رویکرد پیشنهادی کمک کرده و پایه‌ای برای توسعه تحقیقات آینده ایجاد می‌کند.

این پژوهش با نوآوری در استفاده از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی، طراحی یک سیستم یادگیری گروهی برای شناسایی حملات و تمرکز بر حملات تزریق پایگاه‌داده غیررابطه‌ای، گامی مهم در راستای ارتقای امنیت پایگاه‌داده‌های غیررابطه‌ای و توسعه روش‌های پیشرفته در حوزه امنیت سایبری برداشته است.

در این پایان‌نامه، علاوه بر بررسی و تجزیه‌وتحلیل پژوهش‌های مرتبط، یک روش برای شناسایی و اجرای تزریق پایگاه‌داده غیررابطه‌ای به‌منظور بهبود عملکرد برنامه‌های کاربردی وب پیشنهاد می‌شود. نوآوری‌های اعمال شده از اجرای این راهکار شامل موارد زیر است:

* تشکیل یک مجموعه‌داده شامل ورودی‌های سالم و مخرب و تدوین آن، به‌نحوی‌که این مجموعه‌داده‌ها برای آموزش مدل‌های تشخیص ورودی‌های پایگاه‌داده غیررابطه‌ای مبتنی بر یادگیری ماشین استفاده شده است.
* استخراج چندين ویژگی برای طبقه‌بندی بارهای پایگاه‌داده غیررابطه‌ای. از این ویژگی‌ها برای استخراج ویژگی‌های بارهای پایگاه‌داده غیررابطه‌ای استفاده شده است.
* آموزش چند مدل یادگیری و ارزیابی نتایج آن‌ها با دو مجموعه‌داده آزمون برای بررسی دقت مدل در محیط واقعی.

## ساختار فصول پایان‌نامه

فصل دوم: مرور مبانی نظری و پیشینه پژوهش

این فصل به بررسی مفاهیم بنیادی و مرور تحقیقات مرتبط اختصاص دارد. محتوای این فصل شامل موارد زیر است:

* بررسی نقاط ضعف و چالش‌های امنیتی مرتبط با حملات تزریق داده در محیط‌های پایگاه‌داده غیررابطه‌ای.
* مرور مبانی یادگیری ماشین، الگوریتم‌های مرتبط و کاربردهای آن‌ها در حوزه امنیت سایبری.
* معرفی مدل‌های زبان بزرگ و مقایسه آن‌ها با روش‌های سنتی مانند شبکه‌های مولد متخاصم (GAN) برای تولید داده‌های مصنوعی.
* تحلیل و مرور پژوهش‌های پیشین در زمینه تشخیص حملات تزریق پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای.
* بررسی دستاوردهای کلیدی تحقیقات مرتبط و تحلیل نقاط قوت و محدودیت‌های آن‌ها.
* بیان نقاط تمایز رویکرد پیشنهادی این پژوهش نسبت به کارهای پیشین.

فصل سوم: روش پیشنهادی

این فصل به تشریح روش پیشنهادی برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای اختصاص دارد. موضوعات مطرح‌شده در این فصل عبارت‌اند از:

* معرفی و توصیف رویکرد کلی پژوهش.
* شرح فرایند تولید داده‌های مصنوعی با استفاده از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ)، شامل تعریف ویژگی‌های موردنیاز و مراحل تولید مجموعه‌داده.
* معرفی مدل‌های یادگیری ماشینی انتخاب‌شده، شامل Logistic Regression، Random Forest، و XGBoost، و دلایل استفاده از آن‌ها.
* ارائه جزئیات مربوط به تنظیم پارامترها و روش‌های بهینه‌سازی مدل‌ها.
* توضیح معماری کلی سیستم پیشنهادی و مراحل پیاده‌سازی آن.
* بررسی چالش‌ها و نحوه رفع آن‌ها در روش پیشنهادی.

فصل چهارم: پیاده‌سازی و ارزیابی

در این فصل، فرایند پیاده‌سازی و ارزیابی روش پیشنهادی به‌تفصیل شرح داده می‌شود. این فصل شامل موارد زیر است:

* شرح فرایند آماده‌سازی مجموعه‌داده‌ها، شامل داده‌های اولیه و مصنوعی.
* پیاده‌سازی مدل‌ها و تنظیمات موردنیاز برای اجرای آن‌ها.
* تشریح معیارهای ارزیابی مدل‌ها، شامل صحت، نرخ مثبت کاذب[[58]](#footnote-59)، زمان پردازش و قابلیت تعمیم‌پذیری.
* مقایسه عملکرد روش پیشنهادی با روش‌های موجود در زمینه تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای.

فصل پنجم: تحلیل نتایج

این فصل به تحلیل و ارائه نتایج نهایی اختصاص دارد. محتوای این فصل شامل موارد زیر است:

* بررسی نتایج به‌دست‌آمده از آموزش و ارزیابی مدل‌های پیشنهادی.
* تحلیل مزایا و معایب روش پیشنهادی در مقایسه با روش‌های پیشین.
* بررسی قابلیت‌های روش پیشنهادی در مواجهه با داده‌های ناشناخته.
* تحلیل تأثیر استفاده از مدل‌های زبان بزرگ بر کیفیت و کارایی مجموعه‌داده و مدل‌های یادگیری ماشینی.

فصل ششم: جمع‌بندی و کارهای آتی

در این فصل، به جمع‌بندی کلی پژوهش پرداخته شده و دستاوردهای اصلی آن بیان می‌شود. همچنین:

* نقاط قوت و نوآوری‌های پژوهش مرور می‌شود.
* محدودیت‌های پژوهش شناسایی و تشریح می‌گردد.
* پیشنهاداتی برای تحقیقات آتی در زمینه امنیت سایبری و بهبود روش‌های یادگیری ماشینی ارائه می‌شود.

# مبانی نظری و پیشینه پژوهش

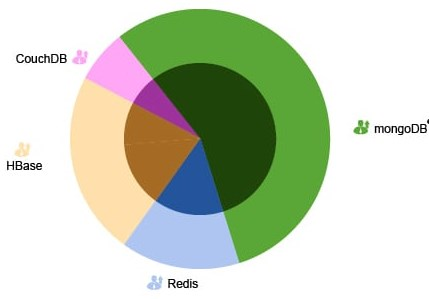
در این فصل به بررسی مفاهیم بنیادی و پژوهش‌های مرتبط با این موضوع اختصاص دارد. ابتدا نقاط ضعف و چالش‌های امنیتی مرتبط با حملات تزریق داده در محیط‌های پایگاه‌داده غیررابطه‌ای بررسی می‌شود. سپس، مبانی یادگیری ماشینی، الگوریتم‌های مرتبط، و کاربردهای آن‌ها در امنیت سایبری موردبحث قرار می‌گیرد. در ادامه، مدل‌های زبان بزرگ به‌عنوان رویکردی نوین برای تولید داده‌های مصنوعی معرفی و با روش‌های سنتی مانند شبکه‌های مولد متخاصم مقایسه می‌شوند. همچنین، تحقیقات پیشین در زمینه تشخیص حملات تزریق پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای مرور شده و دستاوردهای کلیدی آن‌ها تحلیل می‌شود. درنهایت، محدودیت‌های پژوهش‌های پیشین و نقاط تمایز رویکرد پیشنهادی این پژوهش نسبت به آن‌ها بیان می‌گردد.

## 2-1 مبانی نظری

### شناخت پایگاه‌داده‌های غیررابطه‌ای

در سال‌های اخیر، استفاده از فناوری اطلاعات به طور چشمگیری افزایش‌یافته و تولید داده‌های عظیمی را به دنبال داشته است. بسیاری از شرکت‌ها برای مدیریت این داده‌ها از پایگاه‌داده‌های پایگاه‌داده غیررابطه‌ای و روش‌های دفاعی مبتنی بر جاوا اسکریپت استفاده می‌کنند. MongoDB از جمله پایگاه‌داده‌های پایگاه‌داده غیررابطه‌ای است که قدرتمند و ایمن است[28].

پايگاه‌داده Mongodb به دليل فراگيرتر بودن در توسعه نرم‌افزارها هدف اصلي پياده‌سازي ماست. ازآنجايي‌که مجموعه‌داده عمدتاً از طريق خراش دادن وب[[59]](#footnote-60) و داده‌هاي افزوده به دست مي‌آيد، هدف خاص آموزش مدلي متناسب با پايگاه‌داده MongoDB را برآورده مي‌کند. ترکيب مجموعه‌داده براي دستيابي به اهداف موردنظر در اين مقاله کافي است.MongoDB محبوب‌ترين پایگاه‌داده غیررابطه‌ای است که در بين 500 سازمان برتر جهان مورد پذيرش قابل‌توجهي قرار گرفته است [27].



شکل 2- 1: محبوبيت در بين 500شركت برتر جهان[27]

MONGODB امکانات CURD را فراهم می‌کند و سیستم پایگاه‌داده امکان جستجو و ویرایش اطلاعات موجود در آن را برای کاربران فراهم می‌کند. تزریق مخرب می‌تواند منجر به مشکلات امنیتی جدی شود و هکرها می‌توانند کدهای مخرب را با تزریق به ورودی اجرا کنند تا به اطلاعات پایگاه‌داده دسترسی پیدا کنند.

|  |
| --- |
| پایگاه‌داده رابطه‌ای:  “select \* from student where (phone = ‘ “ + ph\_number +” ’);”  MongoDB :  db.collection.find ({phone= ph\_number}) |

شکل 2- 2: تفاوت كوئري پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای

MongoDB ارائه‌دهنده مقیاس‌پذیری بالا، کارایی عالی و امکانات موردنیاز برای مقیاس‌پذیری است. این پایگاه‌داده از فرمت‌های JSON و BSON پشتیبانی می‌کند و توانایی مدیریت میلیون‌ها یا حتی میلیاردها رکورد را دارا است. از "فهرست‏گذاری بر روي رشته‏ها"[[60]](#footnote-61) و شی‌ء گذاری[[61]](#footnote-62) پشتیبانی می‌کند و به‌صورت زمان واقعی عمل می‌کند. همچنین زبان پرس‌وجو MongoDB به نام Mongo است.

### بررسی نقاط ضعف و چالش‌های امنیتی مرتبط با حملات تزریق داده در محیط‌های پایگاه‌داده غیررابطه‌ای

پایگاه‌داده غیررابطه‌ای به‌عنوان یکی از نوآوری‌های برجسته در حوزه مدیریت داده‌ها، برای پاسخگویی به نیازهای پیچیده و متغیر برنامه‌های مدرن توسعه یافته‌اند. این پایگاه‌ها به دلیل انعطاف‌پذیری بالا، مقیاس‌پذیری افقی، و توانایی مدیریت داده‌های ساختاریافته و غیرساختاریافته، به طور گسترده در سیستم‌های بزرگ نظیر شبکه‌های اجتماعی، تجارت الکترونیک، و اینترنت اشیا استفاده می‌شوند. بااین‌حال، این ویژگی‌های برجسته، همراه با نبود استانداردهای امنیتی یکپارچه و معماری غیرسنتی، باعث ظهور نقاط ضعف و چالش‌های امنیتی خاصی در برابر حملات سایبری، به‌ویژه حملات تزریق داده، شده‌اند.

یکی از مهم‌ترین ویژگی‌های پایگاه‌داده‌های غیررابطه‌ای، ساختار منعطف آن‌ها است که برخلاف پایگاه‌داده‌های رابطه‌ای پایگاه‌داده رابطه‌ای، از مدل‌های داده متنوعی مانند مستند محور، گرافی، کلید - مقدار و جدولی پشتیبانی می‌کند. این انعطاف‌پذیری اگرچه مزایای قابل‌توجهی دارد، اما به دلیل نبود یک استاندارد امنیتی جامع، بستری برای سوءاستفاده مهاجمان فراهم می‌کند. بسیاری از پایگاه‌داده‌های غیررابطه‌ای بدون کنترل دسترسی قوی یا ابزارهای احراز هویت اولیه عرضه می‌شوند که می‌تواند منجر به نفوذ و تزریق داده‌های مخرب شود[29].

در پایگاه‌داده‌های رابطه‌ای، زبان پرس‌وجوی ساختاریافته پایگاه‌داده رابطه‌ای یک استاندارد شناخته‌شده است که قابلیت‌های امنیتی نظیر استفاده از پارامترهای آماده را ارائه می‌دهد. در مقابل، پایگاه‌داده‌های غیررابطه‌ای از زبان‌های پرس‌وجوی اختصاصی استفاده می‌کنند که معمولاً فاقد سازوکارهای مشابه برای جلوگیری از تزریق داده هستند. این عدم استانداردسازی باعث افزایش پیچیدگی در تشخیص و جلوگیری از حملات تزریق پایگاه‌داده غیررابطه‌ای می‌شود.

بسیاری از پایگاه‌داده‌های غیررابطه‌ای از سیاست‌های کنترل دسترسی ساده یا ناکافی استفاده می‌کنند. این موضوع، به‌ویژه در مواردی که سیستم‌ها برای کارایی بیشتر بهینه شده‌اند، می‌تواند منجر به دسترسی غیرمجاز به داده‌های حساس شود. برای مثال، در برخی از پایگاه‌داده‌های غیررابطه‌ای، احراز هویت به‌صورت پیش‌فرض غیرفعال است یا تنظیمات امنیتی به طور کامل به توسعه‌دهندگان واگذار می‌شود.

ابزارهای امنیتی پیشرفته که برای پایگاه‌داده‌های رابطه‌ای توسعه یافته‌اند، اغلب با پایگاه‌داده‌های غیررابطه‌ای ناسازگار هستند یا به‌صورت محدودی عمل می‌کنند. این کمبود ابزارهای امنیتی و نظارتی، فرایند شناسایی حملات تزریق و سایر تهدیدات امنیتی را دشوارتر می‌سازد. علاوه بر این، نبود ابزارهای مانیتورینگ مناسب می‌تواند باعث شود که حملات سایبری برای مدت طولانی بدون شناسایی باقی بمانند[29].

اگرچه پایگاه‌داده‌های غیررابطه‌ای به‌طورکلی برای مقیاس‌پذیری طراحی شده‌اند، اما تضمین امنیت در مقیاس بالا یک چالش اساسی است. با افزایش تعداد گره‌ها و حجم داده‌ها، مدیریت امنیت به‌صورت یکپارچه دشوارتر می‌شود و ممکن است نقاط ضعف جدیدی در سیستم ایجاد کند.

حملات تزریق پایگاه‌داده غیررابطه‌ای به دلیل تنوع ساختار داده‌ها و زبان‌های پرس‌وجو، اغلب به‌صورت الگوهای پیچیده و غیرمعمول ظاهر می‌شوند. این امر باعث می‌شود شناسایی این الگوها با استفاده از روش‌های سنتی امنیت سایبری دشوار باشد. برای مقابله با این چالش، استفاده از روش‌های مبتنی بر یادگیری ماشینی که توانایی تحلیل الگوهای رفتاری و داده‌ای را دارند، ضروری است.

پایگاه‌داده‌های غیررابطه‌ای به دلیل توانایی مدیریت داده‌های غیرساختاریافته، به طور گسترده در اینترنت اشیا استفاده می‌شوند. اما ترکیب حجم بالای داده‌های تولیدشده در اینترنت اشیا و نبود سیاست‌های امنیتی قوی در پایگاه‌داده‌های غیررابطه‌ای، یک نقطه‌ضعف عمده ایجاد می‌کند. مهاجمان می‌توانند از این ضعف‌ها برای حملات تزریق و دسترسی غیرمجاز به داده‌های حساس دستگاه‌های اینترنت اشیا استفاده کنند[30].

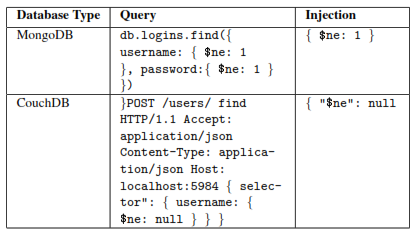
یکی از موانع اصلی در مقابله با حملات تزریق پایگاه‌داده غیررابطه‌ای، نبود مجموعه‌داده‌های استاندارد و قابل‌اعتماد برای آموزش و ارزیابی مدل‌های یادگیری ماشینی است. بسیاری از پژوهش‌ها بر روی داده‌های ساختگی یا مجموعه‌داده‌های محدود انجام می‌شوند که ممکن است قابلیت تعمیم‌پذیری کافی نداشته باشند.

نقاط ضعف و چالش‌های امنیتی پایگاه‌داده‌های غیررابطه‌ای نشان می‌دهد که این سیستم‌ها اگرچه از لحاظ عملکرد و مقیاس‌پذیری مزایای زیادی دارند، اما در برابر حملات تزریق داده آسیب‌پذیر هستند. رفع این چالش‌ها نیازمند ترکیبی از اقدامات امنیتی نظیر اعمال سیاست‌های کنترل دسترسی قوی، توسعه ابزارهای نظارتی مناسب، و استفاده از روش‌های پیشرفته یادگیری ماشینی برای شناسایی و پیشگیری از حملات است. در این راستا، تولید مجموعه‌داده‌های استاندارد و متنوع می‌تواند به‌عنوان یکی از گام‌های اساسی در تحقیقات امنیتی پایگاه‌داده غیررابطه‌ای مطرح باشد.

تکنیک‌هایی که برای تزریق به پایگاه‌داده‌های پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای استفاده می‌شوند، تا حد زیادی مشابه هستند. در اینجا چهار نوع از تزریق‌های پایگاه‌داده غیررابطه‌ای توضیح داده شده‌اند:

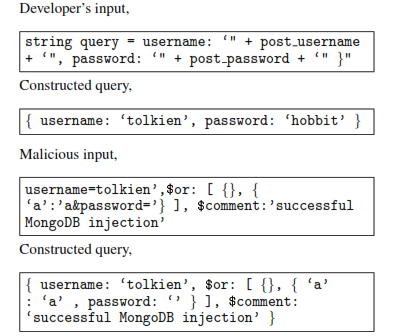
* تزریق تئولوژی‌های[[62]](#footnote-63) : مانند حملات تزریق پایگاه‌داده رابطه‌ای، تزریق پایگاه‌داده غیررابطه‌ای نیز امکان دورزدن فرایند احراز هویت را با تزریق کد در عبارات شرطی فراهم می‌کند و عبارت‌هایی تولید می‌کند که همیشه مقدار true دارند[31]. برای مثال:

جدول 2- 1: انواع دیتا بیس ها



این پرس‌وجوها اطلاعاتی را نشان می‌دهند که در آن نام کاربری و رمز عبور NULL نیستند. هکرها می‌توانند از عملگر "$ne" (به معنای "نابرابر") استفاده کنند تا بدون واردکردن نام کاربری و رمز عبور صحیح به سیستم وارد شوند[30].

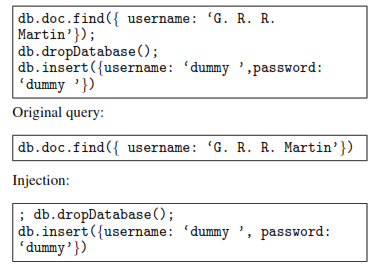
* پرس‌وجوهای Union: هکر از یک پارامتر آسیب‌پذیر برای تغییر داده‌هایی که قرار بود از یک پرس‌وجو بازگردانده شود، استفاده می‌کند. یک شرط OR برای اتصال یک عبارت خالی به ورودی استفاده می‌شود. ازآنجاکه یک عبارت خالی همیشه معتبر است، بررسی رمز عبور بی‌اثر می‌شود. به‌عنوان‌مثال:



در اینجا پرس‌وجوی خالی {} همیشه مقدار true دارد [

* تزریق جاوا اسکریپت[[63]](#footnote-64) : پایگاه‌های داده پایگاه‌داده غیررابطه‌ای اجازه اجرای کدهای جاوا اسکریپت را می‌دهند و می‌توانند پرس‌وجوها و تراکنش‌های پیچیده‌ای را روی موتور پایگاه‌داده اجرا کنند. اگر ورودی کاربر فیلتر یا اعتبارسنجی نشود، ممکن است خطر تزریق کدهای تصادفی جاوا اسکریپت وجود داشته باشد.
* پرس‌وجوهای زنجیره‌ای [[64]](#footnote-65) : در این روش، هکر از توالی‌های فرار[[65]](#footnote-66) و کاراکترهای ویژه‌ای مانند بازگشت به خط [[66]](#footnote-67)، خط جدید [[67]](#footnote-68)، آکولادهای بسته، و علامت نقطه‌ویرگول برای پایان‌دادن به یک پرس‌وجو و سپس افزودن پرس‌وجوهای مخرب اضافی برای اجرا استفاده می‌کند. این عمل می‌تواند پایگاه‌داده را به‌شدت خراب کند. برای مثال:

پرس‌وجو:



در اینجا، پس از یک نقطه‌ویرگول (;)، هکر یک پرس‌وجوی مخرب اضافی را تزریق می‌کند[30].نقض مبدا[[68]](#footnote-69) : رابط برنامه‌نویسی REST در HTTP، نوع جدیدی از آسیب‌پذیری‌ها را به وجود آورده است که به مهاجم اجازه می‌دهد حتی از یک دامنه دیگر به پایگاه‌داده غیررابطه‌ای حمله کند. در حملات Cross-Origin، یک کاربر مجاز و مرورگر وب او مورد سوءاستفاده قرار می‌گیرند تا هکر بتواند عملیات نامطلوبی را اجرا کند. به شکل جعل درخواست بین‌سایتی ، این حمله زمانی رخ می‌دهد که اعتماد سایتی به مرورگر کاربر مورد سوءاستفاده قرار می‌گیرد تا یک عملیات غیرقانونی روی پایگاه‌داده غیررابطه‌ای اجرا شود. با تزریق یک فرم HTیادگیری ماشین در یک وب‌سایت آسیب‌پذیر یا فریب کاربر برای بازدید از وب‌سایت شخصی هکر، هکر می‌تواند یک عملیات پست را روی پایگاه‌داده اجرا کند.

### یادگیری ماشین

یادگیری ماشین یکی از شاخه‌های برجسته هوش مصنوعی است که به ماشین‌ها امکان می‌دهد از داده‌ها یاد بگیرند و بدون نیاز به برنامه‌نویسی صریح، عملکرد خود را بهبود دهند. در حوزه امنیت سایبری، یادگیری ماشین نقش مهمی در شناسایی تهدیدات، پیشگیری از حملات، و تحلیل رفتارهای غیرعادی ایفا می‌کند. این بخش به‌مرور مفاهیم بنیادین یادگیری ماشین، بررسی الگوریتم‌های مرتبط، و کاربردهای آن‌ها در امنیت سایبری اختصاص دارد.

یادگیری ماشین فرایندی است که در آن سیستم‌های کامپیوتری توانایی یادگیری از داده‌ها و بهبود عملکرد خود در انجام وظایف خاص را پیدا می‌کنند. این فرایند شامل مراحل زیر است[32]:

* ورود داده‌ها: داده‌ها به‌عنوان ورودی به سیستم داده می‌شوند.
* استخراج ویژگی‌ها: ویژگی‌های مرتبط با مسئله از داده‌ها استخراج می‌شوند.
* مدل‌سازی: مدل‌های آماری یا الگوریتم‌های یادگیری برای شناسایی الگوها استفاده می‌شوند.
* پیش‌بینی: مدل، داده‌های جدید را تحلیل کرده و خروجی پیش‌بینی می‌کند.
* ارزیابی: عملکرد مدل بر اساس معیارهایی نظیر دقت، یادآوری و نرخ مثبت کاذب ارزیابی می‌شود.

یادگیری ماشین به سه دسته اصلی تقسیم می‌شود:

1. یادگیری نظارت شده[[69]](#footnote-70) :

در این روش، مدل‌ها از داده‌های برچسب‌گذاری شده آموزش می‌بینند. این رویکرد برای مسائل دسته‌بندی و پیش‌بینی مناسب است.

* + مثال: تشخیص حملات سایبری بر اساس الگوهای مشخص حملات پیشین.

1. یادگیری بدون نظارت: [[70]](#footnote-71)

مدل‌ها در این روش از داده‌های بدون برچسب استفاده می‌کنند و به کشف ساختارها و الگوهای مخفی در داده‌ها می‌پردازند.

* + مثال: شناسایی خوشه‌های مشکوک در ترافیک شبکه.

1. یادگیری تقویتی[[71]](#footnote-72) :

این روش مبتنی بر یادگیری از طریق تعامل با محیط است، به‌طوری‌که عامل هوشمند با دریافت پاداش یا تنبیه به یادگیری می‌پردازد.

* + مثال: طراحی سیستم‌های تطبیقی برای شناسایی حملات در زمان واقعی.

#### 2-1-4-1 الگوریتم‌های مرتبط در یادگیری ماشین

الگوریتم‌های پایه‌ای[33]

1. رگرسیون لجستیک

این الگوریتم برای مسائل دسته‌بندی دودویی استفاده می‌شود و احتمال وقوع یک رویداد را مدل‌سازی می‌کند.

* + کاربرد: شناسایی بسته‌های شبکه‌ای مخرب.

1. درخت تصمیم[[72]](#footnote-73)

این مدل بر اساس گره‌هایی که معیارهای تصمیم‌گیری خاصی را بررسی می‌کنند، عمل می‌کند.

* + کاربرد: تحلیل رفتار کاربران و شناسایی دسترسی‌های غیرمجاز.

1. ماشین بردار پشتیبان

الگوریتمی است که داده‌ها را با استفاده از یک ابر صفحه به دودسته جدا می‌کند.

* + کاربرد: تشخیص بدافزار با تفکیک ویژگی‌های بدافزار از نرم‌افزارهای سالم.

1. االگوریتم (KNN[[73]](#footnote-74)) یکی از ساده‌ترین و در عین حال قدرتمندترین الگوریتم‌های یادگیری ماشین است که در دسته‌ی یادگیری نظارت‌شده قرار می‌گیرد. این الگوریتم برای طبقه‌بندی و رگرسیون به کار می‌رود، اما عمدتاً در مسائل طبقه‌بندی مورد استفاده قرار می‌گیرد.

الگوریتم‌های پیشرفته[34]

1. جنگل تصادفی

یک الگوریتم یادگیری گروهی که از ترکیب چندین درخت تصمیم برای افزایش دقت و کاهش بیش برازش استفاده می‌کند.

* + کاربرد: تحلیل داده‌های بزرگ امنیتی.

1. XGBoost:

یکی از قدرتمندترین الگوریتم‌های گروهی است که برای مسائل طبقه‌بندی و رگرسیون استفاده می‌شود.

* + کاربرد: شناسایی حملات پیچیده در ترافیک شبکه.

1. یادگیری عمیق:

این روش بر معماری‌های شبکه عصبی عمیق تکیه دارد و توانایی درک داده‌های پیچیده؛ مانند تصاویر، صدا، و متون را دارد.

* + کاربرد: شناسایی بدافزارهای ناشناخته.

در جدول زیر، مقایسه‌ای دقیق از مدل‌های رگرسیون لجستیک، جنگل تصادفی و XGBoost بر اساس معیارهای مختلف ارائه شده است:

جدول3- 1: مقایسه‌ای دقیق از مدل‌ها

|  |  |  |  |
| --- | --- | --- | --- |
| معیار | Logistic Regression | Random Forest | XGBoost |
| نوع الگوریتم | مدل خطی | یادگیری گروهی | تقویت گرادیان |
| پیچیدگی محاسباتی | پایین | متوسط | بالا |
| تعامل با داده‌های غیرخطی | ضعیف | قوی | بسیار قوی |
| مقاومت در برابر بیش برازش | متوسط | بالا | بسیار بالا |
| سرعت پردازش | سریع | متوسط | سریع |
| تفسیرپذیری مدل | بسیار ساده | متوسط | پیچیده |
| مدیریت داده‌های نامتوازن | ضعیف | متوسط | قوی |
| نیاز به تنظیم پارامترها | کم | متوسط | زیاد |
| دقت در داده‌های پیچیده | پایین | بالا | بسیار بالا |

در ادامه کاربرد یادگیری ماشین در امنیت سایبری را شرح می‌دهیم[35]:

* کاربردهای تشخیص نفوذ: یکی از مهم‌ترین کاربردهای یادگیری ماشین، شناسایی نفوذ و فعالیت‌های غیرمجاز در سیستم‌های کامپیوتری است. الگوریتم‌ها با تحلیل ترافیک شبکه می‌توانند رفتارهای مشکوک را شناسایی و اقدامات لازم را انجام دهند.
* شناسایی بدافزار: الگوریتم‌های یادگیری ماشینی می‌توانند با تحلیل ویژگی‌های فایل‌ها و کدها، بدافزارها را شناسایی کنند. به‌عنوان‌مثال، Random Forest و XGBoost برای شناسایی بدافزارهای پیچیده بسیار مؤثر هستند.
* تحلیل لاگ‌ها: سیستم‌های امنیتی از یادگیری ماشینی برای تحلیل حجم عظیمی از لاگ‌های تولیدشده توسط شبکه‌ها و سیستم‌ها استفاده می‌کنند. این تحلیل به شناسایی الگوهای مخرب کمک می‌کند.
* تشخیص حملات تزریق: الگوریتم‌های نظارت شده نظیر Logistic Regression و SVM برای تشخیص حملات تزریق پایگاه‌داده رابطه‌ای و پایگاه‌داده غیررابطه‌ای با تحلیل داده‌های ورودی و الگوهای مخرب استفاده می‌شوند.
* پیشگیری از حملات فیشینگ: با استفاده از یادگیری ماشینی، ایمیل‌ها و صفحات وب مشکوک بررسی و الگوهای حملات فیشینگ شناسایی می‌شوند. الگوریتم‌های نظارت شده برای این منظور بسیار کاربردی هستند.
* شناسایی رفتارهای مشکوک کاربران: الگوریتم‌های بدون نظارت، مانند خوشه‌بندی، می‌توانند رفتارهای کاربران را مدل‌سازی کرده و رفتارهای غیرعادی را شناسایی کنند.
* سیستم‌های توصیه‌گر امنیتی: با تحلیل الگوهای تهدید و آسیب‌پذیری، مدل‌های یادگیری ماشینی می‌توانند پیشنهادهایی برای ارتقا امنیت سیستم‌ها ارائه دهند.

چالش‌ها و محدودیت‌های یادگیری ماشین در امنیت سایبری عبارت‌اند از:

* کمبود مجموعه‌داده‌های جامع: بسیاری از مدل‌های یادگیری ماشینی برای عملکرد بهینه به داده‌های بزرگ و متنوع نیاز دارند. در حوزه امنیت سایبری، مجموعه‌داده‌های استاندارد و متنوع به‌سختی قابل‌دسترس هستند.
* حملات به مدل‌های یادگیری ماشینی: مهاجمان ممکن است مدل‌های یادگیری ماشینی را هدف حملات خود قرار دهند. برای مثال:
* حملات تزریق داده مخرب: تزریق داده‌های جعلی برای تغییر عملکرد مدل.
* حملات استنتاج مدل: شناسایی رفتارهای مدل برای استفاده از نقاط ضعف آن.
* تغییرات در الگوهای تهدید: الگوهای حملات سایبری به‌سرعت تغییر می‌کنند که باعث می‌شود مدل‌های یادگیری ماشینی نیاز به به‌روزرسانی مداوم داشته باشند.

یادگیری ماشینی به‌عنوان یکی از ابزارهای کلیدی در امنیت سایبری، توانایی تحلیل و شناسایی تهدیدات پیچیده را فراهم می‌کند. از تشخیص بدافزارها گرفته تا شناسایی رفتارهای مشکوک کاربران، این فناوری نقش مهمی در ایمن‌سازی سیستم‌ها ایفا می‌کند. بااین‌حال، چالش‌هایی نظیر کمبود داده‌های استاندارد و تغییرات سریع در الگوهای تهدیدات نشان می‌دهد که توسعه و پیاده‌سازی راهکارهای یادگیری ماشینی در امنیت سایبری نیازمند تحقیق و تلاش مداوم است.

در این پژوهش، تلاش می‌شود با استفاده از مدل‌های پیشرفته نظیر Random Forest و XGBoost و تولید مجموعه‌داده‌های مصنوعی با مدل‌های زبان بزرگ، این چالش‌ها تاحدامکان مرتفع شوند و راهکاری جامع برای مقابله با حملات سایبری ارائه شود.

### معرفی مدل‌های زبان بزرگ و مقایسه آن‌ها با روش‌های سنتی درتولید داده‌های مصنوعی

مدل‌های زبان بزرگ و شبکه‌های مولد متخاصم هر دو به‌عنوان ابزارهای پیشرفته در تولید داده‌های مصنوعی مطرح هستند. این مدل‌ها، هرچند در اهداف و کاربردها مشترکاتی دارند، اما از لحاظ ساختار، روش کار و قابلیت‌ها تفاوت‌های چشمگیری نشان می‌دهند. در این بخش، ابتدا مدل‌های زبان بزرگ معرفی می‌شوند و سپس مقایسه‌ای جامع میان این مدل‌ها و روش‌های سنتی مانند شبکه‌های مولد متخاصم ارائه می‌شود.[36]

مدل‌های زبان بزرگ، معماری‌هایی مبتنی بر شبکه‌های عصبی ترانسفورمر[[74]](#footnote-75)هستند که برای پردازش زبان طبیعی طراحی شده‌اند. این مدل‌ها به دلیل آموزش بر روی مقادیر عظیمی از داده‌های متنی، قابلیت درک و تولید متون پیچیده را دارا هستند. ویژگی‌های کلیدی مدل‌های زبان بزرگ عبارت‌اند از:

* معماری ترانسفورمر[37]:
* این مدل‌ها از معماری ترانسفورمر بهره می‌برند که امکان پردازش داده‌ها به‌صورت موازی و یادگیری روابط طولانی‌مدت میان کلمات و جملات را فراهم می‌کند.
* مقیاس‌پذیری بالا
* مدل‌های زبان بزرگ‌ها بر روی مجموعه‌داده‌های عظیم و با استفاده از میلیاردها پارامتر آموزش‌دیده‌اند. به‌عنوان‌مثال، GPT-3 با 175 میلیارد پارامتر [38]یکی از پیشرفته‌ترین مدل‌های زبان بزرگ محسوب می‌شود.
* توانایی تولید داده‌های متنی واقعی:
* مدل‌های زبان بزرگ می‌توانند متن‌هایی تولید کنند که از نظر ساختار و معنا به متون واقعی شباهت زیادی دارند.
* کاربردهای مدل‌های زبان بزرگ در تولید داده‌های مصنوعی
* شبیه‌سازی داده‌های متنی:
* مدل‌های زبان بزرگ‌ها قادرند داده‌های متنی مصنوعی تولید کنند که می‌تواند شامل پرس‌وجوهای پایگاه‌داده، پیام‌های کاربران، یا داده‌های امنیتی باشد.
* تولید داده‌های متنوع برای امنیت سایبری:
* این مدل‌ها می‌توانند داده‌های مرتبط با حملات سایبری مانند تزریق پایگاه‌داده غیررابطه‌ای را باکیفیت بالا و تنوع زیاد تولید کنند.
* ایجاد داده‌های ترکیبی:

مدل‌های زبان بزرگ‌ها می‌توانند داده‌هایی ترکیبی شامل متن و کد تولید کنند که برای آموزش مدل‌های یادگیری ماشینی در حوزه امنیت سایبری بسیار مفید است.

شبکه‌های مولد متخاصم معماری‌هایی متشکل از دو شبکه عصبی هستند:

* مولد[[75]](#footnote-76): وظیفه تولید داده‌های مصنوعی را دارد.
* تفکیک‌کننده[[76]](#footnote-77) : وظیفه تشخیص داده‌های مصنوعی از داده‌های واقعی را برعهده دارد.

این دو شبکه در رقابتی متقابل عمل می‌کنند، به‌طوری‌که مولد تلاش می‌کند داده‌هایی تولید کند که نتوانند توسط تفکیک‌کننده شناسایی شوند. ویژگی‌های کلیدی شبکه‌های مولد متخاصم عبارت‌اند از:

* تخصص در داده‌های غیرمتنی:

شبکه‌های مولد متخاصم در تولید داده‌های تصویری و سیگنال‌های پیچیده نظیر صدا یا ویدئو بسیار موفق عمل کرده‌اند.

* عملکرد مبتنی بر رقابت:

مولد و تفکیک‌کننده در فرایندی رقابتی بهبود پیدا می‌کنند که باعث افزایش کیفیت داده‌های تولیدشده می‌شود.

محدودیت‌های شبکه‌های مولد متخاصم را در ادامه بیان می‌کنیم:

* محدودیت در تولید داده‌های متنی پیچیده:

معماری شبکه‌های مولد متخاصم به طور خاص برای داده‌های متنی طراحی نشده است و در شبیه‌سازی داده‌های متنی که نیاز به درک زبان طبیعی دارد، محدودیت‌هایی دارد.

* چالش در همگرایی:

فرایند آموزش شبکه‌های مولد متخاصم گاهی دچار ناپایداری می‌شود و ممکن است مولد نتواند داده‌هایی باکیفیت تولید کند.

در ادامه این بخش مقایسه مدل‌های زبان بزرگ و شبکه‌های مولد متخاصم برای تولید داده‌های مصنوعی می‌پردازیم[39], [40]:

**1**-تفاوت در ساختار

* مدل‌های زبان بزرگ بر معماری ترانسفورمر و پردازش زبان طبیعی تمرکز دارد و برای تولید داده‌های متنی بسیار مناسب است.
* شبکه‌های مولد متخاصم برای داده‌های غیرمتنی (تصاویر، سیگنال‌ها و غیره) طراحی شده و بر رقابت میان مولد و تفکیک‌کننده استوار است.

2- توانایی در تولید داده‌های متنی

* مدل‌های زبان بزرگ: به‌دلیل تمرکز بر پردازش زبان، داده‌های متنی پیچیده و شبیه به واقعیت تولید می‌کند.
* شبکه‌های مولد متخاصم: توانایی کمتری در تولید داده‌های متنی با ساختار پیچیده دارد.

3- کاربرد در امنیت سایبری

* مدل‌های زبان بزرگ می‌تواند داده‌های مصنوعی مرتبط با حملات سایبری، پرس‌وجوهای تزریق پایگاه‌داده غیررابطه‌ای و متون مرتبط با امنیت را تولید کند.
* شبکه‌های مولد متخاصم بیشتر برای تولید داده‌های تصویری یا داده‌هایی که تحلیل بصری نیاز دارند (مانند تصاویر نقشه‌برداری بدافزارها) کاربرد دارد.

4- سهولت پیاده‌سازی

* مدل‌های زبان بزرگ تنظیم و استفاده از مدل‌های زبان بزرگ‌ها برای تولید داده‌های متنی ساده‌تر است.
* شبکه‌های مولد متخاصم فرایند آموزش پیچیده‌تر و مستلزم تنظیمات دقیق برای رسیدن به داده‌های باکیفیت است.

5- چالش‌های فنی

* مدل‌های زبان بزرگ نیازمند منابع محاسباتی قوی و داده‌های گسترده برای آموزش است.
* شبکه‌های مولد متخاصم مستعد مشکلاتی نظیر همگرایی ناپایدار یا تولید داده‌های بی‌کیفیت در صورت عدم تنظیم دقیق است.

مدل‌های زبان بزرگ به دلایل زیر برای تولید داده‌های مصنوعی مرتبط با حملات پایگاه‌داده غیررابطه‌ای نسبت به شبکه‌های مولد متخاصم برتری دارند:

1. توانایی درک ساختار پرس‌وجوهای پایگاه‌داده:

مدل‌های زبان بزرگ‌ها می‌توانند پرس‌وجوهای پیچیده و شبه کدهای مرتبط با حملات پایگاه‌داده غیررابطه‌ای را تولید کنند.

1. تولید داده‌های متنی با تنوع بالا:

برخلاف شبکه‌های مولد متخاصم، مدل‌های زبان بزرگ قادرند داده‌های متنی با ساختارها و الگوهای متنوع تولید کنند که برای امنیت سایبری حیاتی است.

1. سادگی در تولید داده‌های ترکیبی:

مدل‌های زبان بزرگ‌ها می‌توانند داده‌هایی ترکیبی از متن و کد تولید کنند که برای شبیه‌سازی حملات سایبری و آموزش مدل‌ها ایده‌آل است.

چالش‌ها و محدودیت‌های مدل‌های زبان بزرگ و GAN

1-چالش‌های مدل‌های زبان بزرگ

* هزینه بالای محاسباتی برای آموزش.
* نیاز به تنظیمات دقیق برای تولید داده‌های متناسب با اهداف خاص.

2- چالش‌های GAN

* ناپایداری در فرایند آموزش.
* محدودیت در تولید داده‌های متنی.

مدل‌های زبان بزرگ به دلیل قابلیت درک زبان طبیعی، توانایی تولید داده‌های متنی پیچیده و انعطاف‌پذیری بالا، جایگزینی مناسب و پیشرفته برای روش‌های سنتی مانند GAN در تولید داده‌های مصنوعی هستند[39]. در حوزه امنیت سایبری و به‌ویژه تولید داده‌های مرتبط با حملات پایگاه‌داده غیررابطه‌ای، استفاده از مدل‌های زبان بزرگ می‌تواند بهبود قابل‌توجهی در کیفیت و تنوع داده‌ها ایجاد کند. این ویژگی‌ها باعث می‌شود که مدل‌های زبان بزرگ‌ها ابزاری ارزشمند در تحقیقات امنیت سایبری و آموزش مدل‌های یادگیری ماشینی باشند.

## تزریق پایگاه‌داده غیررابطه‌ای

در ادامه چند پژوهش انجام شده مورد بررسی قرار میگیرند :

نام مقالهSECURE-D: Framework For Detecting and Preventing Attacks in SQL and NoSQL Databases [41]

این مقاله یک چارچوب امنیتی تحت عنوان SECURE-D را معرفی می‌کند که برای شناسایی و جلوگیری از حملات به پایگاه داده‌های SQL و NoSQL طراحی شده است. این چارچوب از طریق یک پراکسی معکوس عمل می‌کند و تمامی درخواست‌های HTTP را قبل از رسیدن به سرور برنامه مورد بررسی قرار می‌دهد. این سیستم می‌تواند حملات SQL Injection و NoSQL Injection را شناسایی کند و در صورت شناسایی تهدید، از اجرای درخواست‌های مخرب جلوگیری نماید.  
بررسی‌ها نشان می‌دهند که SECURE-D در برابر حملات مختلف مانند تزریق SQL (مثل Tautologies، Union Queries، Piggy-backed Queries، و Stored Procedures) و همچنین حملات NoSQL (مثل NoSQL JavaScript Injection و Cross-Origin Violations) مؤثر است.

مهم‌ترین مشارکت‌های این مقاله عبارت‌اند از:

* ارائه یک چارچوب جدید برای محافظت هم‌زمان از پایگاه‌های داده SQL و NoSQL در برابر حملات تزریقی، که این امر در تحقیقات پیشین مورد بررسی قرار نگرفته بود.
* استفاده از پراکسی معکوس برای تجزیه و تحلیل درخواست‌های ورودی پیش از رسیدن به سرور، که باعث افزایش امنیت وب‌سایت‌ها در برابر تهدیدات اینترنتی می‌شود.
* شناسایی انواع مختلف حملات SQL و NoSQL و جلوگیری از اجرای درخواست‌های آلوده در سطح پایگاه داده.
* ثبت و نمایش گزارشات حملات برای مدیران امنیتی، همراه با قابلیت مسدودسازی دستی آدرس‌های IP مخرب.
* مقایسه با روش‌های موجود مانند SEA WAF و DNIARS، که نشان می‌دهد SECURE-D طیف وسیع‌تری از حملات را شناسایی و مسدود می‌کند.

متدولوژی مقاله: محققان برای پیاده‌سازی و ارزیابی SECURE-D، سه ماژول اصلی را در معماری سیستم معرفی کرده‌اند:

1. ماژول تجزیه‌گر [[77]](#footnote-78)
   * درخواست‌های HTTP ورودی را تجزیه کرده و به URL، داده‌های GET و POST تبدیل می‌کند.
   * داده‌های ورودی را از نظر وجود الگوهای مخرب بررسی می‌کند.
2. ماژول تشخیص حمله [[78]](#footnote-79)
   * شامل یک لیست از کلیدواژه‌های مخرب SQL و NoSQL مانند "SELECT"، "DROP"، "DELETE" برای SQL و "insertOne"، "create" برای NoSQL است.
   * در صورت شناسایی هر یک از این کلمات کلیدی در ورودی، درخواست مسدود می‌شود.
   * می‌تواند حملاتی مانند Tautologies، Logically Incorrect Queries، Union Queries، Piggy Backed Queries، و Stored Procedures در SQL و NoSQL Injection، NoSQL JavaScript Injection، و Cross-Origin Violations را شناسایی کند.
3. ماژول ثبت لاگ [[79]](#footnote-80)
   * تمامی حملات شناسایی‌شده را ثبت کرده و اطلاعاتی نظیر تعداد حملات، نوع حمله، و آدرس IP منبع حمله را نمایش می‌دهد.
   * مدیران امنیتی می‌توانند به‌صورت دستی IPهای مشکوک را مسدود کنند.

این سیستم در یک وب‌سایت آسیب‌پذیر تست شده است و نشان داده که می‌تواند به‌صورت لحظه‌ای حملات را شناسایی و متوقف کند.

محدودیت‌های مقاله

* محدودیت در شناسایی حملات پیچیده: اگر حمله‌کننده از تکنیک‌های Obfuscation (مبهم‌سازی) برای پنهان کردن کد مخرب خود استفاده کند، ممکن است سیستم SECURE-D در شناسایی آن‌ها دچار مشکل شود.
* نیاز به تنظیمات دستی برای برخی از حملات: این سیستم به‌طور خودکار همه حملات را مسدود نمی‌کند و در برخی موارد نیاز به دخالت دستی مدیران امنیتی برای به‌روزرسانی لیست کلیدواژه‌های مخرب وجود دارد.
* عدم بررسی سایر حملات امنیتی: تمرکز این مقاله صرفاً بر روی حملات تزریقی (SQL و NoSQL) است و حملات دیگری مانند XSS (Cross-Site Scripting) یا CSRF (Cross-Site Request Forgery) را پوشش نمی‌دهد.
* وابستگی به معماری پراکسی معکوس: این سیستم فقط در صورتی کارایی دارد که در مسیر تمامی درخواست‌های ورودی به سرور قرار گیرد، که ممکن است در برخی از زیرساخت‌های ابری یا سیستم‌های توزیع‌شده چالش‌برانگیز باشد.

نتیجه‌گیری مقاله

* SECURE-D یک چارچوب کارآمد برای شناسایی و جلوگیری از حملات پایگاه داده‌های SQL و NoSQL است.
* این سیستم درخواست‌های HTTP ورودی را پردازش کرده و حملات تزریقی را در زمان اجرا شناسایی و مسدود می‌کند.
* بررسی‌های عملی نشان دادند که این روش در مقایسه با روش‌های دیگر مانند SEA WAF و DNIARS عملکرد بهتری دارد و دامنه وسیع‌تری از حملات را پوشش می‌دهد.
* مزیت اصلی این چارچوب، ترکیب قابلیت‌های تشخیص حملات SQL و NoSQL در یک سیستم واحد است، که در روش‌های قبلی وجود نداشت.
* استفاده از ماژول‌های تجزیه و تحلیل، شناسایی و ثبت گزارشات باعث افزایش امنیت پایگاه داده‌های تحت وب می‌شود.

مقاله(Basic NoSQL Injection Analysis and Detection on MongoDB) [42]

این مقاله به بررسی حملات تزریق NoSQL در پایگاه داده MongoDB می‌پردازد. با توجه به افزایش استفاده از پایگاه‌های داده NoSQL به دلیل ویژگی‌هایی مانند مقیاس‌پذیری بالا، انعطاف‌پذیری و دسترسی سریع به داده‌ها، این پایگاه‌ها نیز مانند SQL در معرض حملات تزریقی قرار دارند.

این مطالعه به تحلیل آسیب‌پذیری‌های NoSQL و نحوه اجرای حملات تزریق در MongoDB می‌پردازد. برای مقابله با این تهدیدات، نویسندگان دو روش دفاعی را پیشنهاد می‌کنند:

1. اعتبارسنجی ورودی کاربران
2. استفاده از دستورات پارامتری [[80]](#footnote-81)

نتایج نشان می‌دهند که MongoDB، با وجود مزایای خود، همچنان در برابر حملات تزریقی آسیب‌پذیر است و باید اقدامات امنیتی لازم برای مقابله با این تهدیدات انجام شود.

مشارکت‌های کلیدی این تحقیق شامل موارد زیر است:

* تحلیل آسیب‌پذیری پایگاه داده‌های NoSQL در برابر حملات تزریقی، که نشان می‌دهد با وجود استفاده از JSON Queries، همچنان این پایگاه‌ها در معرض تهدید هستند.
* بررسی نحوه اجرای حملات تزریق در پایگاه داده MongoDB و شبیه‌سازی این حملات با استفاده از زبان‌های PHP و JavaScript.
* ارائه روش‌های مقابله با حملات تزریق NoSQL، شامل:
  1. محدود کردن ورودی کاربران با اعتبارسنجی مقادیر ورودی.
  2. استفاده از دستورات پارامتری به جای درج مستقیم داده‌های ورودی در دستورات پایگاه داده.
* ایجاد یک مدل برای تشخیص حملات تزریقی که به توسعه‌دهندگان کمک می‌کند تا آسیب‌پذیری‌های سیستم خود را شناسایی و برطرف کنند.

این مقاله بر حملات تزریق NoSQL در MongoDB تمرکز دارد و روش‌های اجرای این حملات و همچنین روش‌های مقابله با آن‌ها را مورد بررسی قرار داده است.

1. مراحل تحقیق و آزمایش:
2. تحلیل آسیب‌پذیری‌های MongoDB:
   * بررسی ساختار کلی پایگاه داده MongoDB و عملکرد آن.
   * تجزیه و تحلیل نحوه اجرای حملات تزریقی در محیط NoSQL.
3. اجرای حملات تزریقی در MongoDB:
   * استفاده از PHP و JavaScript برای ارسال کدهای مخرب به پایگاه داده.
   * بررسی آسیب‌پذیری‌های ورودی‌های کاربر، که ممکن است باعث اجرای کدهای مخرب شود.
   * نمایش نمونه‌ای از یک حمله تزریق که از طریق فرم ورودی انجام شده و داده‌های غیرمجاز را از پایگاه داده استخراج می‌کند.
4. پیشنهاد راهکارهای دفاعی:
   * استفاده از اعتبارسنجی ورودی: جلوگیری از ورود مقادیر غیرمجاز در فیلدهای ورودی.
   * استفاده از دستورات پارامتری: جدا کردن ورودی کاربر از دستورات پایگاه داده برای جلوگیری از اجرای کدهای مخرب.
   * استفاده از فیلترهای خاص برای حذف کاراکترهای مخرب در درخواست‌های کاربر.

با وجود ارائه راهکارهای مؤثر برای مقابله با حملات تزریقی NoSQL، این تحقیق دارای چند محدودیت است:

* محدود بودن مطالعه به MongoDB

این مقاله تنها MongoDB را بررسی کرده و روش‌های تزریق NoSQL در سایر پایگاه‌های داده مانند Cassandra، Redis یا CouchDB بررسی نشده‌اند.

* عدم بررسی تکنیک‌های پیچیده‌تر تزریق NoSQL

این مقاله بر حملات پایه‌ای NoSQL تمرکز دارد و روش‌های پیچیده‌تری مانند تزریق چند مرحله‌ای یا ترکیب حملات را پوشش نمی‌دهد.

* عدم ارائه آزمایش عملی در مقیاس وسیع:

این مطالعه بر یک محیط آزمایشی کوچک انجام شده و تأثیر حملات در سیستم‌های واقعی و پیچیده سازمانی مورد بررسی قرار نگرفته است.

* عدم بررسی سایر تهدیدات امنیتی NoSQL

مقاله تنها به حملات تزریقی تمرکز دارد و سایر تهدیدات امنیتی مانند حملات XSS، CSRF و حملات مبتنی بر رمزنگاری را پوشش نمی‌دهد.

نتیجه‌گیری مقاله

* MongoDB و سایر پایگاه‌های داده NoSQL همچنان در معرض حملات تزریقی قرار دارند، اگرچه از SQL استفاده نمی‌کنند.
* روش‌های تزریق در NoSQL شباهت زیادی به SQL Injection دارد، اما به جای SQL، از JSON Query Language برای اجرا استفاده می‌شود.
* حملات تزریقی در NoSQL می‌توانند منجر به دسترسی غیرمجاز به اطلاعات، حذف داده‌ها و حتی تخریب کامل پایگاه داده شوند.
* برای افزایش امنیت، توسعه‌دهندگان باید از روش‌هایی مانند اعتبارسنجی ورودی، استفاده از دستورات پارامتری و فیلتر کردن مقادیر کاربر استفاده کنند.
* استفاده از یک مدل تشخیص حمله، می‌تواند به شناسایی فعالیت‌های مخرب کمک کند و امنیت پایگاه داده را افزایش دهد.

مقاله NoSQL Racket: A Testing Tool for Detecting NoSQL Injection Attacks in Web Applications [43]

این مقاله یک ابزار جدید به نام NoSQL Racket را معرفی می‌کند که برای شناسایی حملات تزریق NoSQL در برنامه‌های وب طراحی شده است. حملات تزریق NoSQL به این دلیل رخ می‌دهند که برنامه‌های وب، ورودی‌های کاربر را بدون اعتبارسنجی به پرس‌وجوهای پایگاه داده وارد می‌کنند. این مسئله به مهاجمان اجازه می‌دهد که کدهای مخرب خود را به پایگاه داده ارسال کنند و از این طریق به اطلاعات حساس دسترسی پیدا کنند یا آن‌ها را تغییر دهند.

ویژگی اصلی ابزار NoSQL Racket این است که ساختار کدهای پرس‌وجوی NoSQL را در دو سطح تحلیل استاتیک (در کد منبع) و تحلیل دینامیک (در زمان اجرا) بررسی و با یکدیگر مقایسه می‌کند. این ابزار روی چهار نوع پایگاه داده مختلف (MongoDB، Cassandra، CouchDB و Amazon DynamoDB) آزمایش شده و اثبات شده که توانایی شناسایی حملات تزریق NoSQL را دارد، در حالی که ابزارهای تست رایج مانند Netsparker، Vega و Skipfish در این زمینه ناکام بوده‌اند.

مهم‌ترین مشارکت‌های این تحقیق شامل موارد زیر است:

1. ارائه یک ابزار جدید برای شناسایی حملات تزریق NoSQL که برخلاف ابزارهای رایج، بدون نیاز به زبان مشخص و مدل داده‌ای خاص، قادر به شناسایی حملات NoSQL در پایگاه‌های داده مختلف است.
2. بررسی و تحلیل روش‌های مختلف تزریق NoSQL در پایگاه‌های داده MongoDB، Cassandra، CouchDB و Amazon DynamoDB.
3. مقایسه ابزار NoSQL Racket با ابزارهای تست امنیت رایج (Netsparker، Vega و Skipfish) و نشان دادن برتری آن در تشخیص حملات تزریقی NoSQL.
4. ارائه یک مدل ترکیبی برای شناسایی حملات NoSQL که ترکیبی از تحلیل استاتیک کد و تحلیل دینامیک اجرای پرس‌وجو را به کار می‌گیرد.
5. اجرای آزمایش‌های عملکردی برای بررسی سرعت و کارایی NoSQL Racket در محیط‌هایی با بارگذاری‌های سنگین و کاربرهای هم‌زمان.

مراحل اصلی متدولوژی تحقیق و پیاده‌سازی:

1. تحلیل حملات تزریق NoSQL:
   * بررسی انواع حملات تزریق NoSQL که از طریق فرم‌های ورودی، کوکی‌ها و درخواست‌های HTTP انجام می‌شوند.
   * ارائه یک مثال عملی از نحوه‌ی تزریق NoSQL در MongoDB که به مهاجم اجازه می‌دهد بدون داشتن رمز عبور معتبر، وارد سیستم شود.
2. طراحی و پیاده‌سازی ابزار NoSQL Racket:
   * این ابزار از جدولی به نام "Driverstbl" استفاده می‌کند که شامل کلمات کلیدی، عملگرهای منطقی و عملگرهای رابطه‌ای NoSQL است.
   * کدهای پرس‌وجوی NoSQL در دو حالت (تحلیل استاتیک و دینامیک) مورد بررسی قرار می‌گیرند و ساختار آن‌ها با یکدیگر مقایسه می‌شود.
   * در صورتی که ساختار پرس‌وجو در زمان اجرا با ساختار اولیه مطابقت نداشته باشد، ابزار حمله را تشخیص داده و اجرای درخواست را متوقف می‌کند.
3. مقایسه با سایر ابزارهای امنیتی:
   * ابزار NoSQL Racket روی چهار پایگاه داده مختلف آزمایش شده است.
   * ابزارهای تست امنیتی معروف Netsparker، Vega و Skipfish به کار گرفته شدند اما نتوانستند حملات NoSQL را شناسایی کنند.
   * ابزار NoSQL Racket موفق شد حملات را در تمامی پایگاه‌های داده آزمایش‌شده تشخیص دهد.
4. آزمایش‌های عملکردی:
   * ابزار NoSQL Racket با استفاده از ابزار LoadComplete تحت بارگذاری سنگین قرار گرفت.
   * نتایج نشان دادند که این ابزار می‌تواند ۵۰ کاربر هم‌زمان در هر ثانیه را مدیریت کند.
   * میانگین زمان بارگذاری صفحات ۷۵ میلی‌ثانیه بود که نشان‌دهنده‌ی عملکرد مناسب ابزار است.

با وجود اثربخشی ابزار NoSQL Racket، این تحقیق دارای برخی محدودیت‌ها است:

* عدم پوشش تمام پایگاه‌های داده NoSQL

ابزار فقط روی MongoDB، Cassandra، CouchDB و Amazon DynamoDB آزمایش شده و سایر پایگاه‌های داده مانند Redis، Neo4j و Firebase بررسی نشده‌اند.

* محدودیت در شناسایی حملات پیچیده‌تر:

ابزار فقط حملات تزریق مستقیم را شناسایی می‌کند و ممکن است در برابر تزریقات پیشرفته‌تر مانند حملات چندمرحله‌ای (Multi-step Injection) یا ترکیبی با XSS و CSRF دچار مشکل شود.

* وابستگی به تنظیمات اولیه:

ابزار NoSQL Racket نیاز دارد که ابتدا جدول Driverstbl را به درستی تنظیم کند. در صورتی که یک پایگاه داده جدید معرفی شود، باید این جدول به‌روزرسانی شود.

* عدم بررسی سایر تهدیدات امنیتی:

مقاله تنها به حملات تزریق NoSQL پرداخته و سایر حملات امنیتی مانند سرقت نشست‌ها، XSS و حملات مبتنی بر رمزنگاری را پوشش نداده است.

نتیجه‌گیری مقاله

* NoSQL Injection یک تهدید جدی برای امنیت پایگاه‌های داده NoSQL است و بسیاری از برنامه‌های وب در برابر این نوع حملات آسیب‌پذیر هستند.
* ابزارهای تست امنیت رایج مانند Netsparker، Vega و Skipfish قادر به تشخیص حملات NoSQL نیستند.
* ابزار NoSQL Racket می‌تواند به طور دقیق حملات تزریقی NoSQL را شناسایی کند و مانع از اجرای پرس‌وجوهای مخرب شود.
* این ابزار بر روی چهار پایگاه داده‌ی NoSQL مختلف آزمایش شده و نتایج نشان داده‌اند که در تمامی موارد حملات را شناسایی کرده است.
* آزمایش‌های عملکردی نشان دادند که این ابزار تحت بارگذاری‌های سنگین عملکرد خوبی دارد و می‌تواند در محیط‌های واقعی مورد استفاده قرار گیرد.

مقاله (Analysis on Database Security Model Against NoSQL Injection)[44]

این مقاله به بررسی حملات تزریق NoSQL و راهکارهای امنیتی برای مقابله با این تهدیدات در پایگاه‌های داده NoSQL مانند MongoDB و Cassandra می‌پردازد. با توجه به افزایش استفاده از پایگاه‌های داده NoSQL به دلیل مقیاس‌پذیری بالا، انعطاف‌پذیری و دسترسی سریع به داده‌ها، این سیستم‌ها در برابر حملات مختلف از جمله تزریق NoSQL و حملات Cross-Site Request Forgery (CSRF) آسیب‌پذیر شده‌اند.

مدل امنیتی پیشنهادی در این مقاله شامل استفاده از الگوی امنیتی متمرکز بر داده‌ها (Data-Centric Security Model) است که اطلاعات را قبل از ذخیره در پایگاه داده رمزگذاری می‌کند. علاوه بر این، احراز هویت کاربران در شبکه‌های غیرقابل اعتماد از طریق پروتکل Kerberos انجام می‌شود.

* نکات کلیدی مقاله:
* بررسی حملات تزریق NoSQL و نحوه سوءاستفاده مهاجمان از آن‌ها.
* بررسی آسیب‌پذیری‌های MongoDB و Cassandra در برابر حملات تزریقی.
* معرفی پروتکل امنیتی Kerberos برای احراز هویت کاربران و افزایش امنیت پایگاه داده.
* بررسی روش‌های جلوگیری از حملات، از جمله اعتبارسنجی ورودی کاربران و رمزگذاری داده‌ها.
* آزمایش حملات تزریق NoSQL با استفاده از JavaScript و PHP.

مهم‌ترین مشارکت‌های این تحقیق شامل موارد زیر است:

1. تحلیل حملات تزریق NoSQL و ارائه الگویی برای طبقه‌بندی آسیب‌پذیری‌های امنیتی در این پایگاه‌های داده.
2. بررسی و مقایسه امنیت MongoDB و Cassandra در برابر حملات تزریقی، از جمله تحلیل مکانیزم‌های احراز هویت و رمزگذاری داده‌ها.
3. ارائه یک مدل امنیتی مبتنی بر داده[[81]](#footnote-82) که داده‌ها را قبل از ذخیره شدن در پایگاه داده رمزگذاری می‌کند.
4. استفاده از پروتکل Kerberos برای احراز هویت کاربران در پایگاه داده‌های NoSQL، که می‌تواند به جلوگیری از دسترسی‌های غیرمجاز کمک کند.
5. ارائه راهکارهای امنیتی برای مقابله با حملات تزریق NoSQL، از جمله:
   * اعتبارسنجی ورودی کاربران
   * استفاده از احراز هویت
   * فعال‌سازی ثبت گزارش‌های امنیتی
   * رمزگذاری داده‌ها قبل از ذخیره‌سازی در پایگاه داده

مراحل اصلی تحقیق و پیاده‌سازی:

* تحلیل حملات تزریق NoSQL:
  + بررسی نحوه عملکرد پایگاه‌های داده NoSQL مانند MongoDB و Cassandra.
  + شناسایی نقاط ضعف امنیتی و نحوه اجرای حملات تزریق NoSQL در این سیستم‌ها.
  + ارائه مثال‌هایی از تزریق NoSQL در محیط MongoDB و Node.js.
* بررسی آسیب‌پذیری‌های NoSQL در برابر حملات تزریقی:
  + تحلیل نمونه‌هایی از حملات تزریق در MongoDB که در آن‌ها از کوئری‌های مخرب استفاده شده است.
  + بررسی تأثیر حملات تزریقی در Cassandra و راه‌های جلوگیری از آن.
* ارائه راهکارهای امنیتی برای مقابله با این حملات:
  + معرفی پروتکل Kerberos برای احراز هویت و جلوگیری از دسترسی غیرمجاز.
  + بررسی روش‌های رمزگذاری داده‌ها قبل از ذخیره در پایگاه داده.
  + بررسی روش‌های جلوگیری از اجرای کوئری‌های مخرب در پایگاه داده.
* مقایسه امنیت MongoDB و Cassandra:
  + بررسی مکانیزم‌های امنیتی هر دو پایگاه داده.
  + مقایسه وضعیت احراز هویت، رمزگذاری و محافظت در برابر حملات تزریقی در MongoDB و Cassandra.
* ارائه یک مدل امنیتی ترکیبی:
  + استفاده از Kerberos برای احراز هویت و مدیریت نشست‌های کاربری.
  + رمزگذاری داده‌ها قبل از ذخیره‌سازی برای جلوگیری از حملات تزریقی.
  + استفاده از ثبت گزارش‌های امنیتی (Auditing) برای پایش فعالیت‌های کاربران.

با وجود ارائه راهکارهای امنیتی مؤثر، این تحقیق دارای برخی محدودیت‌ها است:

* تمرکز بر MongoDB و Cassandra

این مطالعه فقط دو پایگاه داده NoSQL را بررسی کرده و سایر پایگاه‌های داده مانند CouchDB، Firebase و Redis پوشش داده نشده‌اند.

* عدم بررسی حملات ترکیبی

تحقیق تنها به حملات تزریق NoSQL پرداخته و سایر تهدیدات مانند XSS، CSRF و حملات مهندسی اجتماعی بررسی نشده‌اند.

* وابستگی به Kerberos

پیشنهاد اصلی مقاله استفاده از Kerberos برای احراز هویت کاربران است، اما این روش در برخی سیستم‌ها به دلیل پیچیدگی پیاده‌سازی دشواری‌هایی دارد.

* عدم آزمایش در محیط‌های واقعی و بزرگ:

آزمایش‌های امنیتی انجام‌شده در یک محیط آزمایشی کوچک صورت گرفته‌اند و تأثیر روش‌های پیشنهادی در سیستم‌های مقیاس‌پذیر سازمانی بررسی نشده است.

نتیجه‌گیری مقاله

* حملات تزریق NoSQL یک تهدید جدی برای امنیت پایگاه‌های داده NoSQL هستند، زیرا این سیستم‌ها به دلیل نداشتن ساختار ثابت، در برابر کوئری‌های مخرب آسیب‌پذیرتر هستند.
* MongoDB و Cassandra به دلیل انعطاف‌پذیری بالا، در برابر حملات تزریقی آسیب‌پذیر هستند.
* پروتکل Kerberos یک راهکار مؤثر برای احراز هویت کاربران و جلوگیری از دسترسی غیرمجاز به پایگاه داده است.
* رمزگذاری داده‌ها قبل از ذخیره در پایگاه داده، یک لایه امنیتی اضافی ایجاد می‌کند و مانع از افشای اطلاعات حساس در صورت وقوع حملات می‌شود.
* فعال‌سازی ثبت گزارش‌های امنیتی (Auditing) می‌تواند به شناسایی حملات و فعالیت‌های مخرب کمک کند.
* پایگاه‌های داده NoSQL نیاز به راهکارهای امنیتی قوی‌تری دارند تا بتوانند در برابر حملات پیچیده‌ای مانند تزریق NoSQL مقاومت کنند.

جدول 2- 2: مقایسه مقالات [42], [43], [44]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **نام مقاله** | **هدف مطالعه** | **روش‌شناسی** | **یافته‌های اصلی** | **شکاف‌های تحقیقاتی** | **نتیجه‌گیری** |
| Basic NoSQL Injection Analysis and Detection on MongoDB[42] | بررسی و تحلیل حملات تزریق NoSQL در پایگاه داده MongoDB و ارائه راهکارهای دفاعی | اجرای حملات تزریق NoSQL با استفاده از PHP و JavaScript و پیشنهاد روش‌های امنیتی برای جلوگیری از حملات | نشان داده شد که پایگاه داده MongoDB در برابر حملات NoSQL Injection آسیب‌پذیر است و می‌توان از روش‌هایی مانند اعتبارسنجی ورودی و استفاده از دستورات پارامتری برای جلوگیری از این حملات استفاده کرد | مطالعه فقط بر روی MongoDB انجام شده و سایر پایگاه‌های داده NoSQL بررسی نشده‌اند | استفاده از اعتبارسنجی ورودی و دستورات پارامتری می‌تواند از حملات تزریق NoSQL جلوگیری کند، اما برای مقابله با تهدیدات پیچیده‌تر نیاز به تحقیقات بیشتر است |
| NoSQL Racket: A Testing Tool for Detecting NoSQL Injection Attacks in Web Applications[43] | طراحی و توسعه یک ابزار امنیتی برای شناسایی و جلوگیری از حملات تزریق NoSQL | ارائه ابزار NoSQL Racket که پرس‌وجوهای NoSQL را در دو سطح تحلیل استاتیک و دینامیک مقایسه کرده و حملات تزریق را شناسایی می‌کند | ابزار NoSQL Racket توانست حملات تزریق NoSQL را با موفقیت شناسایی کند، در حالی که ابزارهای تست امنیتی رایج مانند Netsparker، Vega و Skipfish نتوانستند این حملات را تشخیص دهند | این تحقیق تنها بر روی چهار پایگاه داده (MongoDB، Cassandra، CouchDB و Amazon DynamoDB) متمرکز است و سایر پایگاه‌های داده NoSQL مورد بررسی قرار نگرفته‌اند | ابزار NoSQL Racket روشی کارآمد برای شناسایی حملات تزریق NoSQL است و باید برای پایگاه‌های داده بیشتری آزمایش شود |
| Analysis on Database Security Model Against NoSQL Injection[44] | تحلیل مدل‌های امنیتی پایگاه داده در برابر حملات تزریق NoSQL و ارائه راهکارهای امنیتی | مقایسه امنیت MongoDB و Cassandra و استفاده از پروتکل Kerberos برای احراز هویت و جلوگیری از حملات | استفاده از Kerberos می‌تواند از حملات NoSQL Injection جلوگیری کند و رمزگذاری داده‌ها یک لایه امنیتی اضافی برای محافظت از اطلاعات فراهم می‌کند | این مطالعه بر MongoDB و Cassandra متمرکز است و سایر تهدیدات امنیتی مانند XSS، CSRF و حملات مهندسی اجتماعی را پوشش نمی‌دهد | استفاده از Kerberos و رمزگذاری داده‌ها می‌تواند امنیت پایگاه‌های داده NoSQL را افزایش دهد، اما نیاز به تحقیقات بیشتری در مورد سایر پایگاه‌های داده NoSQL و حملات پیشرفته وجود دارد |

|  |  |  |  |
| --- | --- | --- | --- |
| **معیار مقایسه** | **[42]** | **[43]** | **[44]** |
| نوع پژوهش | تحلیلی – آزمایشی | ابزاری – تشخیصی | معماری – امنیتی |
| تمرکز اصلی | نحوه وقوع حمله | شناسایی حمله | جلوگیری معماری‌محور |
| سطح دفاع | سطح Query | سطح ابزار تست | سطح سیستم و احراز هویت |
| کاربرد عملی | متوسط | بسیار بالا | بالا |
| مناسب برای | فهم حمله | تست نفوذ | طراحی امن سیستم |

### رویکرد یادگیری ماشین

شناسایی و پیشگیری از تزریق پایگاه‌داده غیررابطه‌ای با رویکرد یادگیری ماشین

یک مدل یادگیری ماشین برای شناسایی تزریق پایگاه‌داده غیررابطه‌ای با استفاده از یادگیری نظارت شده مبتنی بر ویژگی‌ها توسعه داده شد[45].

در این روش، آن‌ها مجموعه داده‌ای از پرس‌وجوهای سالم و مخرب مربوط به MongoDB ایجاد کردند، زیرا مجموعه داده‌ای برچسب‌گذاری شده برای تزریق پایگاه‌داده غیررابطه‌ای موجود نبود. آن‌ها از منابع مختلفی مانند OWASP، راهنمای MongoDB و غیره استفاده کردند.

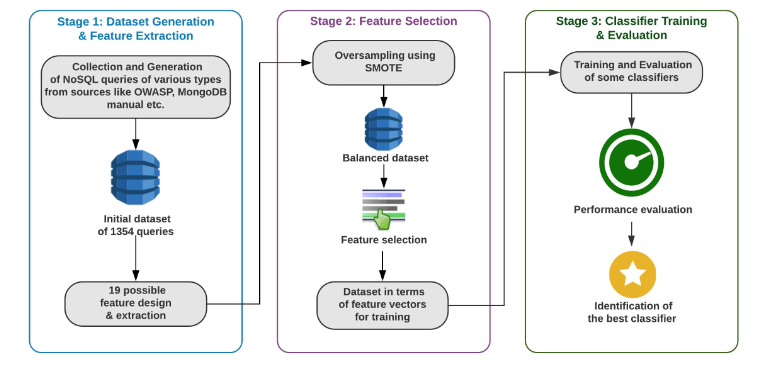
برای انتخاب ویژگی‌های مدل، از ابزار WEKA با روش ClassifierSubsetEval و الگوریتم‌های طبقه‌بندی مانند J48 درخت تصمیم) و IBK (k-نزدیک‌ترین همسایگی و جستجوی گام‌به‌گام حریصانه با حذف معکوس استفاده کردند. این روش‌ها به شناسایی 10 ویژگی برتر بر اساس اطلاعات به‌دست‌آمده[[82]](#footnote-83) و همبستگی[[83]](#footnote-84) کمک کردند[24], [46]. این ویژگی‌های انتخاب‌شده در جدول 2-1 ذکر شده‌اند.

در مجموعه‌داده آن‌ها، انواع مختلفی از حملات مورد بررسی قرار گرفتند، از جمله:

* تزریق آرایه PHP
* تزریق OR در پایگاه‌داده غیررابطه‌ای
* تزریق مبتنی بر جاوا اسکریپت
* پرس‌وجوهای زنجیره‌ای[[84]](#footnote-85)

جدول 2- 3: ویژگی‌های انتخاب‌شده

|  |  |
| --- | --- |
| رتبه | ویژگی |
| 1 | **مقایسه محتویات (Contains Comparison):** بررسی وجود مقادیر مشابه در پرس‌وجو. |
| 2 | **پرس‌وجوی جدید (New Query):** وجود یک پرس‌وجوی جدید در داده‌ها. |
| 3 | **شامل رشته خالی (Contains Empty String):** بررسی حضور رشته‌های خالی در پرس‌وجو. |
| 4 | **شامل عبارت نابرابر (Contains Not Equal):** وجود عملگر "نابرابر" در پرس‌وجو. |
| 5 | **شامل بار مخرب (Contains Payload):** بررسی وجود داده‌های مخرب در پرس‌وجو. |
| 6 | **حضور عبارت بازگشتی (Presence of Return):** بررسی حضور عبارت‌های بازگشتی در پرس‌وجو. |
| 7 | **عبارت همیشه صحیح (Always True Expression):** شناسایی عبارات منطقی که همیشه نتیجه **true** دارند. |
| 8 | **عملیات ارزیابی پرس‌وجو (Evaluation Query Operation):** ارزیابی نحوه عملکرد پرس‌وجوها. |
| 9 | **شامل عملگر منطقی (Contains Logical Operator):** بررسی وجود عملگرهای منطقی مانند **AND** و **OR**. |
| 10 | **عملیات پرس‌وجوی عنصری (Element Query Operation):** انجام عملیات پرس‌وجو بر روی عناصر مشخص‌شده. |



شکل 2- 3: مدل مفهومی یادگیری ماشین[46]

با استفاده از 10 ویژگی ذکر شده در جدول 1-2 در این مطالعه، آن‌ها بر روی دسته‌بندی دودویی کار کردند که شامل دو کلاس سالم و تزریق بود. برای این کار، از طبقه‌بندی‌کننده‌های یادگیری ماشین مانند الگوریتم مبتنی بر درخت تصمیم ID3، شبکه عصبی مصنوعی با پس‌انتشار خطا، جنگل تصادفی، AdaBoost، نزدیک‌ترین همسایه k، ماشین‌های بردار پشتیبان و XGBoost استفاده کردند.

آن‌ها از اعتبارسنجی متقابل 10 قسمتی [[85]](#footnote-86)برای ارزیابی عملکرد طبقه‌بندی‌کننده‌ها استفاده کردند [25]. سپس از یک ابزار تولید تزریق پایگاه‌داده غیررابطه‌ای به نام پایگاه‌داده غیررابطه‌ای Map برای ایجاد مجموعه‌داده تست استفاده کردند [26].

پایگاه‌داده غیررابطه‌ای Map برای تولید مجموعه‌داده اصلی آن‌ها استفاده نشده بود. این مجموعه‌داده تست پر از تزریق پایگاه‌داده غیررابطه‌ای بود و آن‌ها مدل خود را با این مجموعه‌داده آزمایش کردند. روش آن‌ها در مقایسه باSqreen عملکرد بهتری داشت و نرخ شناسایی به طور متوسط 36.25٪ بیشتر از Sqreen بود.

### رویکرد غیر یادگیری ماشین

برخی از روش‌های غیر یادگیری ماشین برای شناسایی و پیشگیری از حملات تزریق پایگاه‌داده غیررابطه‌ای عبارت‌اند از:

1. خودکارساز: در این رویکرد، یک مدل شناسایی مبتنی بر خودکارساز برای تزریق پایگاه‌داده غیررابطه‌ای ایجاد می‌شود [39]. این مدل عمدتاً بر روی حملات تزریق زمان‌بندی‌شده و کور متمرکز است.

در تزریق مبتنی بر زمان، هکر تلاش می‌کند یک تابع جاوا اسکریپت را همراه با یک نشانه معتبر پایگاه‌داده غیررابطه‌ای اضافه کند. با این کار، مهاجم پایگاه‌داده را در حالت تعلیق قرار می‌دهد. برای مثال، پرس‌وجوی زیر، پایگاه‌داده MongoDB را برای 5 ثانیه در حالت تعلیق قرار می‌دهد اگر ورودی آن "John" باشد:

John’,✩where: ‘function(){ sleep(5000); return this.name ==\John’" }

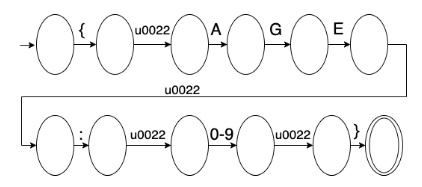
در تزریق کور مبتنی بر بولی (Blind Based Boolean Injection)، مهاجمان از قابلیت‌های MongoDB استفاده می‌کنند تا به لیستی از مجموعه‌ها (Collections)، تعداد مجموعه‌ها و اطلاعات مشابه دسترسی پیدا کنند.

return (db.getCollectionNames().length == 1);

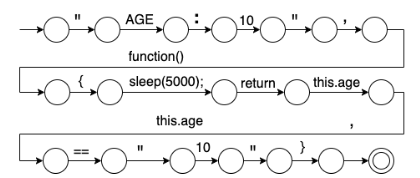
return(tojsononeline (db.collectionname.find() [0]).length == 1);

return(db.getCollectionNames() [0] [0] == ‘a’);

برای شناسایی و پیشگیری از تزریق پایگاه‌داده غیررابطه‌ای، ابتدا نقاطی در کد منبع که از آن‌ها فراخوانی پایگاه‌داده انجام می‌شود، شناسایی می‌کنند. به طور معمول، در MongoDB این نقاط شامل توابع find()، insert()، remove() و update() هستند. سپس برای پرس‌وجوهای معتبر یا ایمن که می‌توانند از این نقاط حساس (Hotspots) تولید شوند، مدل‌های NFA ماشین‌های حالت محدود غیرقطعی) ایجاد می‌کنند و از کتابخانه JSA (تحلیلگر رشته جاوا برای این کار استفاده می‌کنند[47].



شکل 4: نمونه کوئری در منگو دیبی



شکل 2- 4 : مدل نمونه برای ورودی کاربر مورد نظر حمله[47]

سپس آن‌ها پرس‌وجوهای دینامیکی را هنگام دریافت ورودی از کاربر ایجاد می‌کنند. اگر این پرس‌وجوها با مدل‌های NFA ازپیش‌ساخته‌شده مطابقت داشته باشند، پرس‌وجو به‌عنوان معتبر یا ایمن برای سیستم آن‌ها شناخته می‌شود. تنها پرس‌وجوهای معتبر مجاز هستند به پایگاه‌داده ارسال شوند. در شکل 5 می‌توانیم ببینیم که یک ورودی نامعتبر ارسال شده است. اگر شکل 4 و شکل 5 را مقایسه کنیم، متوجه خواهیم شد که مدل‌ها مطابقت ندارند. بنابراین، پرس‌وجوی شکل 4 اجازه ارسال به پایگاه‌داده را نخواهد داشت. به این ترتیب، راه‌حل آن‌ها برای شناسایی و پیشگیری از تزریق پایگاه‌داده غیررابطه‌ای مبتنی بر خودکارساز کار می‌کند.

1. اعتبارسنجی ورودی کاربر: توسعه‌دهندگان در هنگام ساخت سیستم برای شناسایی و پیشگیری از حملات تزریق پایگاه‌داده رابطه‌ای، اقدامات احتیاطی مختلفی انجام می‌دهند. به‌عنوان‌مثال، در MongoDB می‌توان با افزودن کد زیر، اندازه فیلدهای ورودی را محدود کرد[29] :

|  |
| --- |
| onkeypress = return"event.keyCode>= 48&&event.keyCode<=57" |

این کد تنها اعداد را می‌پذیرد. نمادها، فاصله‌ها یا برخی کاراکترهای خاص دیگر نیز بررسی و فیلتر می‌شوند تا از تزریق کد مخرب جلوگیری شود.

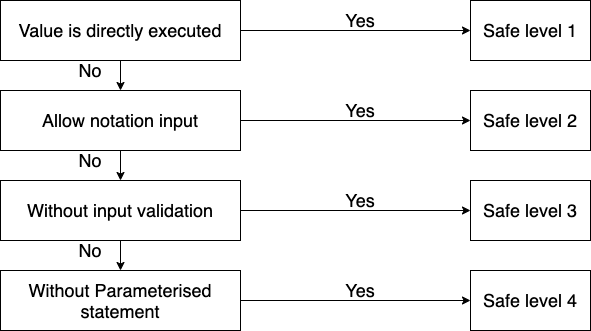
1. پارامتری‌سازی: متغیرهای ورودی کاربر نباید مستقیماً در عبارت شرطی درج شوند و باید فیلتر شوند. در پارامتری‌سازی، از دستورات پارامتری برای ارسال متغیرهای ورودی استفاده می‌شود. به‌جای درج مستقیم متغیرهای ورودی کاربر در عبارت شرطی، از پارامترها استفاده می‌شود. کد مربوطه در زیر نشان‌داده‌شده است:

if(is numeric(✩usearchtwo)=="true"){} else

echo "Incorrect.";

این قطعه کد بررسی می‌کند که آیا پرس‌وجو حاوی عدد است و در صورت وجود، مقدار را می‌پذیرد .[29]

1. شناسایی ویژگی‌های مخرب: شناسایی ویژگی‌های مخرب برای تشخیص این است که آیا سیستم یا نرم‌افزار دارای ویژگی‌هایی است که برای امنیت خطرناک هستند یا خیر. این شناسایی بر اساس کدها و ویژگی‌های مخرب انجام می‌شود. این فرایند می‌تواند به توسعه‌دهندگان کمک کند تا سطح ایمنی پروژه‌های خود را ارزیابی کنند. هرچه عدد سطح ایمنی بالاتر باشد، پایگاه‌داده غیررابطه‌ای امن‌تر است.



شکل 2- 5: فلوچارت کد جاوا اسکریپت برای محدود کردن ورودی[29]

# روش پیشنهادی

این فصل به معرفی و تشریح روش پیشنهادی پژوهش می‌پردازد که هدف آن ارائه راهکاری جامع برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای است. در این روش، از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی باکیفیت و تنوع بالا استفاده می‌شود که می‌تواند محدودیت‌های موجود در دسترسی به مجموعه‌داده‌های واقعی را کاهش دهد. داده‌های تولیدشده در ادامه برای آموزش مدل‌های یادگیری ماشینی مانند Logistic Regression، Random Forest، و XGBoost مورداستفاده قرار می‌گیرند. انتخاب این مدل‌ها بر اساس ویژگی‌های آن‌ها در شناسایی الگوهای پیچیده و قابلیت ترکیب در یک سیستم یادگیری گروهی انجام شده است.

علاوه بر این، روش پیشنهادی شامل تنظیم دقیق پارامترهای مدل‌ها و بهینه‌سازی آن‌ها برای افزایش دقت و کاهش نرخ خطا است. معماری کلی سیستم پیشنهادی و مراحل مختلف پیاده‌سازی، از تولید داده‌های مصنوعی تا ارزیابی مدل‌های یادگیری ماشینی، در این فصل تشریح می‌شود. همچنین، چالش‌های مرتبط با اجرای این روش، از جمله پیچیدگی تولید داده‌ها و تنظیم مدل‌ها، موردبحث قرار گرفته و راه‌حل‌های ارائه‌شده برای غلبه بر آن‌ها بررسی می‌شود.

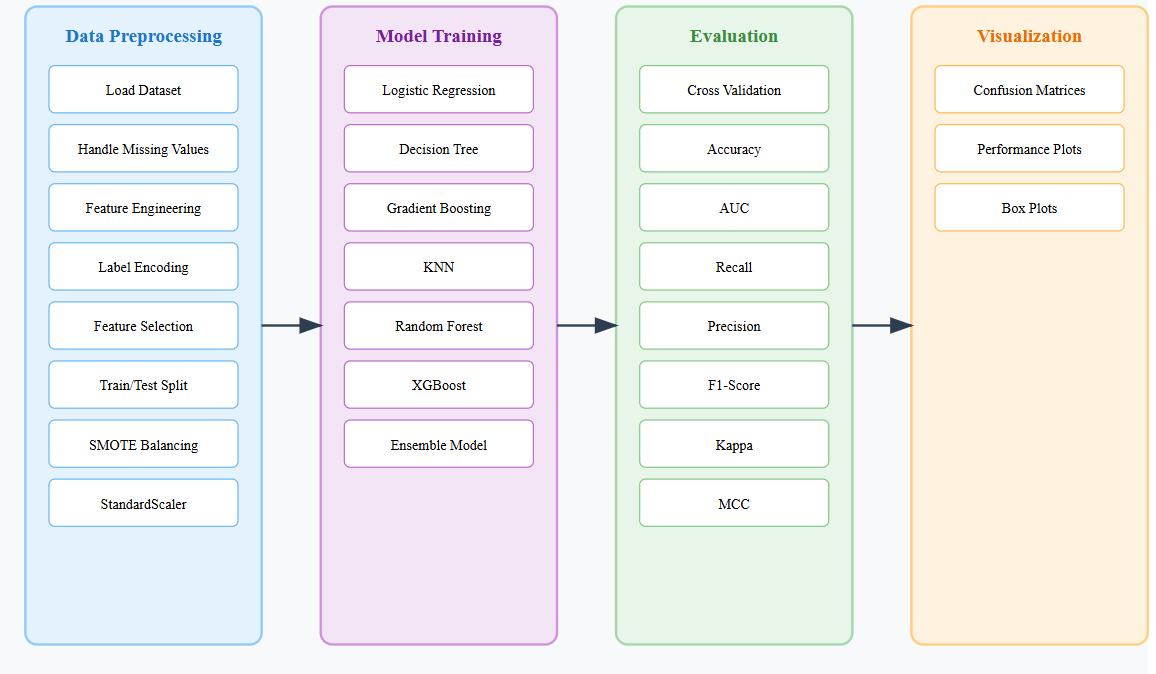
هدف این فصل، ارائه یک نمای جامع از روش پیشنهادی و گام‌های لازم برای پیاده‌سازی آن است تا بتواند به‌عنوان یک ابزار مؤثر در شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای به کار گرفته شود.

* 1. **معرفی و توصیف رویکرد کلی پژوهش**

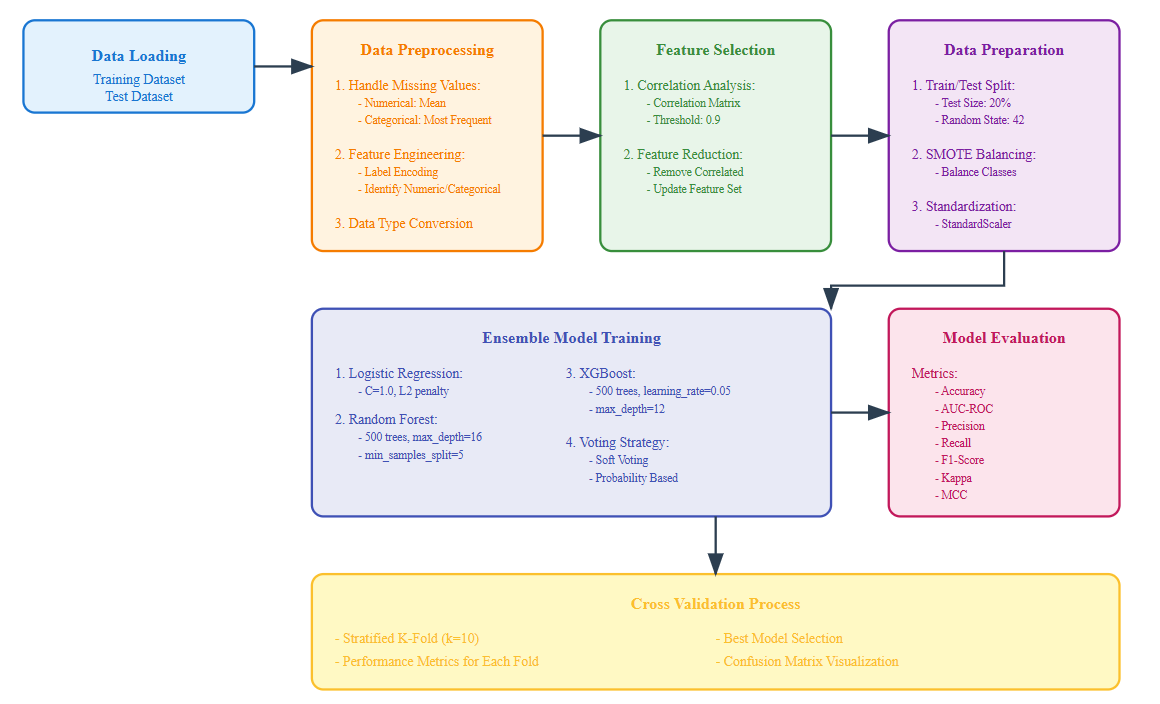
روش پیشنهادی این پژوهش برای مقابله با حملات تزریق پایگاه‌داده غیررابطه‌ای بر پایه استفاده از تکنیک‌های یادگیری ماشینی و تولید داده‌های مصنوعی طراحی شده است. این رویکرد با بهره‌گیری از داده‌های متنوع و الگوریتم‌های یادگیری قوی، تلاش می‌کند تا با شناسایی دقیق الگوهای مخرب، از ورود حملات جلوگیری کند. برای دستیابی به این هدف، رویکرد پژوهش شامل چندین مرحله است که در ادامه تشریح می‌شوند.

اهداف اصلی این پژوهش عبارت‌اند از:

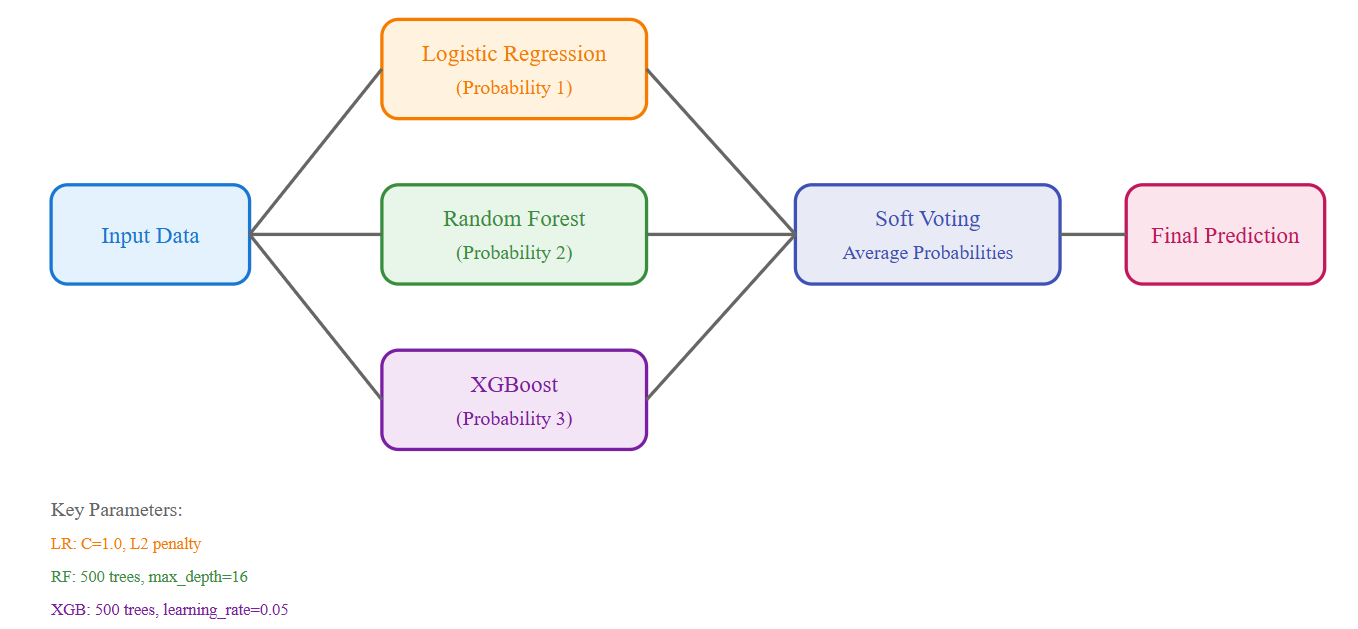
* شناسایی و مقابله با حملات تزریق در پایگاه‌داده‌های غیررابطه‌ای با استفاده از یادگیری ماشینی: هدف این بخش از پژوهش، طراحی و پیاده‌سازی الگوریتم‌های یادگیری ماشینی برای شناسایی و مقابله با حملات تزریق در پایگاه‌داده‌های غیررابطه‌ای است.
* تولید مجموعه‌داده‌های مصنوعی با کیفیت و تنوع بالا برای آموزش مدل‌های یادگیری ماشین: هدف این بخش، ایجاد مجموعه‌داده‌های مصنوعی است که هم از لحاظ کیفیت و هم تنوع، به اندازه کافی برای آموزش و ارزیابی مدل‌های یادگیری ماشینی مؤثر باشند. این مجموعه‌داده‌ها به افزایش دقت شناسایی حملات کمک خواهند کرد.
* افزایش دقت مدل‌های یادگیری ماشینی در شناسایی حملات با استفاده از داده‌های واقعی و مصنوعی: این بخش از پژوهش به دنبال بهبود دقت مدل‌ها با ترکیب داده‌های واقعی و مصنوعی است تا به این ترتیب، مدل‌های یادگیری ماشینی بتوانند شناسایی دقیق‌تری از حملات در پایگاه‌داده‌های غیررابطه‌ای داشته باشند.
* استفاده از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی: یکی از چالش‌های اصلی در زمینه شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای، کمبود مجموعه‌داده‌های مناسب و متنوع است. مجموعه‌داده‌های موجود اغلب محدود به تعداد اندکی از سناریوهای حمله هستند و نمی‌توانند نماینده‌ای جامع از حملات واقعی باشند. برای غلبه بر این چالش، رویکرد پیشنهادی از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ) استفاده می‌کند.



شکل 3- 1: مدل مفهومی کار انجام شده در این پژوهش



شکل 3- 2: مدل مفهومی روش پیشنهادی



شکل 3- 3: دیاگرام روش پیشنهادی

مدل‌های یادگیری ماشینی پیشنهادی

پس از تولید مجموعه‌داده، رویکرد پیشنهادی به سراغ انتخاب مدل‌های یادگیری ماشینی می‌رود. این مدل‌ها به طور خاص برای شناسایی حملات پیچیده طراحی شده‌اند و توانایی بالایی در تحلیل داده‌های ساختاریافته و غیرساختاریافته دارند.

مدل‌های انتخاب‌شده:

1. Logistic Regression:
   * کاربرد: مناسب برای شناسایی الگوهای خطی در داده‌ها و طبقه‌بندی دودویی.
   * دلیل انتخاب: سادگی در پیاده‌سازی و تفسیر، و کارایی بالا در تحلیل اولیه داده‌ها.
2. Random Forest:
   * کاربرد: استفاده از چندین درخت تصمیم برای کاهش خطا و افزایش دقت.
   * دلیل انتخاب: توانایی بالا در تحلیل ویژگی‌های پیچیده و تعمیم‌پذیری بهتر.
3. XGBoost:
   * کاربرد: الگوریتمی پیشرفته برای مسائل طبقه‌بندی و رگرسیون.
   * دلیل انتخاب: کارایی بالا در شناسایی الگوهای پیچیده و سرعت مناسب در پردازش داده‌ها.

یادگیری گروهی:

برای افزایش دقت و کاهش خطا، این پژوهش از ترکیب مدل‌های فوق در یک سیستم یادگیری گروهی استفاده می‌کند. این ترکیب، نقاط قوت هر مدل را با یکدیگر ترکیب کرده و عملکرد کلی را بهبود می‌بخشد.

رویکرد کلی این پژوهش، ترکیبی از تولید داده‌های مصنوعی با استفاده از مدل‌های زبان بزرگ و استفاده از مدل‌های پیشرفته یادگیری ماشینی است. این رویکرد با تمرکز بر تولید داده‌های متنوع و آموزش مدل‌های دقیق، تلاش می‌کند تا حملات تزریق پایگاه‌داده غیررابطه‌ای را بادقت و کارایی بالا شناسایی کند. سیستم پیشنهادی با معماری جامع خود، راهکاری عملی و قابل‌اتکا برای مقابله با این نوع حملات ارائه می‌دهد.

## روش جمع‌آوری داده

در این بخش، فرایند تولید داده‌های مصنوعی برای حملات تزریق پایگاه‌داده غیررابطه‌ای با استفاده از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ) نظیر GPT تشریح می‌شود. این فرایند شامل استخراج ویژگی‌های مهم از مجموعه‌داده اولیه و تولید داده‌های مصنوعی تصادفی و متنوع با استفاده از مدل‌های زبان بزرگ است. در ادامه، مراحل این فرایند به طور علمی و گام‌به‌گام تشریح می‌شود.

**3-3-1 استخراج ویژگی‌های کلیدی از مجموعه‌داده اولیه**

1- مرحله آماده‌سازی مجموعه‌داده اولیه

برای شروع تولید داده‌های مصنوعی، یک مجموعه‌داده اولیه شامل داده‌های مرتبط با پرس‌وجوهای پایگاه‌داده غیررابطه‌ای به مدل GPT ارائه شد. هدف این مرحله شناسایی ویژگی‌های کلیدی از داده‌های واقعی به‌منظور استفاده به‌عنوان مبنایی برای تولید داده‌های مصنوعی بود.

جدول 3- 2: نمونه ای از داده اولیه

|  |  |
| --- | --- |
| text | label |
| {"user": "{}", "password": "{}"} | 0 |
| {"username": {"$regex": "^{}"}} | 1 |

2- درخواست استخراج ویژگی‌ها از مدل مدل‌های زبان بزرگ

مدل GPT با استفاده از یک پرامپت مشخص، وظیفه استخراج ویژگی‌های مهم و معنادار از مجموعه‌داده اولیه را برعهده گرفت. این ویژگی‌ها شامل پارامترهایی بودند که ارتباط مستقیم با ساختار داده‌های پایگاه‌داده غیررابطه‌ای و امکان تشخیص الگوهای مخرب داشتند. نتیجه این مرحله منجر به شناسایی ویژگی‌های زیر شد:

جدول 3- 3: ویژگی های استخراج شده

|  |  |
| --- | --- |
| ویژگی‌ها | شرح ویژگی |
| num\_keys | تعداد کلیدها در ساختار داده. |
| num\_operators | تعداد عملگرهای مورداستفاده در پرس‌وجو. |
| length\_user\_value | طول مقادیر مرتبط با کاربر. |
| length\_password\_value | طول مقادیر مرتبط با رمز عبور. |
| length\_username\_value | طول مقادیر مرتبط با نام کاربری. |
| contains\_email\_pattern | وجود الگوهای آدرس ایمیل در داده‌ها. |
| contains\_ip\_pattern | وجود الگوهای مرتبط با آدرس IP در داده‌ها. |
| nested\_keys\_exist | وجود کلیدهای تودرتو در ساختار داده. |
| num\_nested\_keys | تعداد کلیدهای تودرتو موجود. |
| contains\_numeric\_values | وجود مقادیر عددی در داده‌ها. |
| contains\_boolean\_values | وجود مقادیر بولی (True/False) در داده‌ها. |
| contains\_empty\_strings | وجود مقادیر خالی در داده‌ها. |
| contains\_fixed\_values | وجود مقادیر ثابت (Hardcoded) در داده‌ها. |
| unique\_key\_count | تعداد کلیدهای یکتا در ساختار داده. |
| data\_type\_diversity | تنوع انواع داده‌ها (عددی، رشته‌ای و ...). |
| total\_record\_length | طول کل رکورد داده. |
| special\_characters\_exist | وجود کاراکترهای خاص در مقادیر ورودی. |
| nested\_depth | عمق تودرتو بودن داده‌ها. |
| label | برچسب‌گذاری داده‌ها برای مشخص‌کردن مخرب یا غیرمخرب بودن. |

تحلیل ویژگی‌های استخراج‌شده: ویژگی‌های استخراج‌شده، به‌عنوان معیارهایی برای توصیف رفتار و ساختار داده‌ها به کار گرفته شدند. این ویژگی‌ها نقش کلیدی در تولید داده‌های مصنوعی داشتند، زیرا داده‌های تولیدشده باید این ویژگی‌ها را رعایت کنند تا بتوانند شباهت زیادی به داده‌های واقعی داشته باشند.

* + 1. **تولید داده‌های مصنوعی با استفاده از مدل GPT**

طراحی پرامپت برای تولید داده‌های مصنوعی: پس از استخراج ویژگی‌های کلیدی، در مرحله بعد، از مدل GPT برای تولید داده‌های مصنوعی استفاده شد. برای این کار، پرامپتی طراحی شد که وظیفه تولید داده‌های تصادفی را با رعایت ویژگی‌های مشخص‌شده برعهده داشت. پرامپت شامل موارد زیر بود:

* توصیف ویژگی‌های استخراج‌شده (مانند تعداد کلیدها، عمق تودرتو، وجود مقادیر خاص و ...).
* درخواست تولید داده‌هایی مشابه مجموعه‌داده اولیه با حفظ ساختار و تنوع در ویژگی‌ها.
* تعیین تعداد داده‌های موردنیاز تولید 20.000 نمونه داده مصنوعی.

تولید داده‌های تصادفی: مدل GPT با استفاده از پرامپت تعریف‌شده، داده‌های مصنوعی متنوع و تصادفی تولید کرد. این داده‌ها شامل تمام ویژگی‌های مشخص‌شده در مرحله قبل بودند و از نظر ساختاری و معنایی به داده‌های واقعی شباهت زیادی داشتند. پس از تولید داده‌های مصنوعی، این داده‌ها با مجموعه‌داده واقعی ترکیب شدند تا مجموعه‌داده‌ای جامع و متنوع ایجاد شود. این ترکیب باعث شد تا:

* داده‌های مصنوعی به مدل یادگیری ماشینی کمک کنند تا سناریوهای جدیدی از حملات را یاد بگیرد.
* مدل بتواند تعمیم‌پذیری بالاتری داشته باشد و داده‌های ناشناخته را بهتر شناسایی کند.

دستاوردهای فرایند تولید داده‌های مصنوعی

فرایند تولید داده‌های مصنوعی با استفاده از مدل GPT، نتایج و مزایای زیر را به همراه داشت:

1. ایجاد مجموعه‌داده‌ای جامع و متنوع:

داده‌های تولیدشده شامل سناریوهای مختلف و پیچیده از حملات تزریق پایگاه‌داده غیررابطه‌ای بودند.

1. غلبه بر مشکل کمبود داده:

تولید 20,000 داده مصنوعی، مشکل دسترسی به داده‌های واقعی محدود را برطرف کرد.

1. افزایش کیفیت آموزش مدل‌های یادگیری ماشینی:

داده‌های تولیدشده با تنوع بالا به مدل‌های یادگیری ماشینی کمک کردند تا عملکرد بهتری در تشخیص حملات داشته باشند.

1. صرفه‌جویی در زمان و هزینه:

استفاده از مدل GPT برای تولید داده‌های مصنوعی، فرایند تولید داده را سریع‌تر و مقرون‌به‌صرفه‌تر کرد.

* 1. معرفی مدل‌های Logistic Regression، Random Forest، و XGBoost،

در این بخش، مدل‌های یادگیری ماشینی که در این پژوهش برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای مورداستفاده قرار گرفته‌اند، معرفی و تشریح می‌شوند. این مدل‌ها عبارت‌اند از: رگرسیون لجستیک، جنگل تصادفی، و XGBoost. هر یک از این مدل‌ها به دلیل ویژگی‌ها و قابلیت‌های خاص خود انتخاب شده‌اند که در ادامه به طور دقیق بررسی می‌شوند.

1- رگرسیون لجستیک

رگرسیون لجستیک یک الگوریتم یادگیری نظارت شده است که برای مسائل دسته‌بندی دودویی به کار می‌رود. این الگوریتم با استفاده از تابع سیگموئید احتمال وقوع یک کلاس (مثلاً حمله یا غیر حمله) را محاسبه می‌کند و خروجی را در بازه [0,1] ارائه می‌دهد.

ویژگی‌ها و مزایا

* سادگی و تفسیرپذیری: مدل رگرسیون لجستیک به‌سادگی قابل‌پیاده‌سازی و تفسیر است.
* کارایی بالا در داده‌های کوچک: این مدل برای مجموعه‌داده‌های کوچک و متوسط عملکرد خوبی دارد.
* عدم نیاز به تنظیمات پیچیده: برخلاف مدل‌های پیچیده، رگرسیون لجستیک نیاز به تنظیمات پارامتری زیادی ندارد.
* کاربرد در تشخیص الگوهای خطی: این مدل به‌خوبی می‌تواند داده‌هایی که دارای روابط خطی هستند را تفکیک کند.

دلیل استفاده:

رگرسیون لجستیک به‌عنوان یک مدل پایه برای مقایسه عملکرد سایر مدل‌ها استفاده شده است. این مدل به طور خاص برای تشخیص اولیه الگوهای ساده و خطی در داده‌های مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای مناسب است.

2- جنگل تصادفی

جنگل تصادفی یک الگوریتم یادگیری گروهی است که از ترکیب چندین درخت تصمیم برای انجام فرایند دسته‌بندی یا رگرسیون استفاده می‌کند. هر درخت به‌صورت مستقل آموزش می‌بیند و خروجی نهایی بر اساس رأی‌گیری میان درخت‌ها تعیین می‌شود.

ویژگی‌ها و مزایا

* مقاومت در برابر بیش برازش: به دلیل استفاده از چندین درخت تصمیم، مدل کمتر دچار بیش برازش می‌شود.
* توانایی پردازش داده‌های پیچیده: می‌تواند به‌خوبی الگوهای غیرخطی و پیچیده را شناسایی کند.
* تعامل با ویژگی‌های متنوع: این مدل قادر به تحلیل ویژگی‌های دسته‌ای و عددی به طور هم‌زمان است.
* مقاومت در برابر نویز: در برابر داده‌های نامناسب یا نویز مقاوم است.

دلیل استفاده:

استفاده از جنگل تصادفی به دلیل توانایی بالای آن در تحلیل ویژگی‌های پیچیده و شناسایی الگوهای غیرخطی در داده‌های مربوط به حملات تزریق پایگاه‌داده غیررابطه‌ای ضروری است. این مدل با ترکیب نتایج چندین درخت تصمیم، دقت و کارایی بالایی در تشخیص حملات ارائه می‌دهد.

3- XGBoost

[[86]](#footnote-87)XGBoost یک الگوریتم یادگیری گروهی است که بر اساس تقویت گرادیان عمل می‌کند. در این مدل، درخت‌های تصمیم به‌صورت ترتیبی آموزش می‌بینند و هر درخت تلاش می‌کند خطای درخت قبلی را به حداقل برساند.

ویژگی‌ها و مزایا

* کارایی و سرعت بالا: XGBoost به دلیل بهینه‌سازی‌های داخلی، سرعت بالایی در پردازش داده‌ها دارد.
* کنترل دقیق بیش برازش: این مدل از پارامترهایی مانند Regularization برای جلوگیری از بیش برازش استفاده می‌کند.
* پشتیبانی از داده‌های نامتوازن: XGBoost می‌تواند با تنظیم وزن کلاس‌ها، داده‌های نامتوازن (مانند تعداد کم حملات در مقایسه با داده‌های سالم) را مدیریت کند.
* شناسایی الگوهای پیچیده: این مدل به طور خاص در داده‌هایی که الگوهای غیرخطی دارند، عملکرد بسیار خوبی ارائه می‌دهد.

دلیل استفاده

مدل XGBoost به‌عنوان یک الگوریتم پیشرفته و بهینه‌شده برای حل مسائل طبقه‌بندی پیچیده انتخاب شده است. در این پژوهش، XGBoost به دلیل سرعت بالا و دقت مناسب در شناسایی حملات پیچیده و ناشناخته تزریق پایگاه‌داده غیررابطه‌ای استفاده شده است.

در این پژوهش، سه مدل رگرسیون لجستیک، جنگل تصادفی و XGBoost به‌عنوان الگوریتم‌های اصلی برای شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای انتخاب شدند. این انتخاب بر اساس ویژگی‌های هر مدل و نیاز پژوهش صورت گرفت:

* رگرسیون لجستیک به‌عنوان مدل پایه برای تحلیل اولیه و شناسایی الگوهای خطی.
* جنگل تصادفی به‌عنوان مدلی مقاوم برای شناسایی الگوهای پیچیده بادقت بالا.
* XGBoost به‌عنوان یک مدل پیشرفته و بهینه برای افزایش دقت در تشخیص حملات پیچیده و غیرخطی.

استفاده از این سه مدل به‌صورت ترکیبی در سیستم پیشنهادی، دقت، کارایی و مقاومت مدل را در برابر چالش‌های داده‌های نامتوازن و پیچیده افزایش می‌دهد.

**3-5 ارائه جزئیات مربوط به تنظیم پارامترها**

در این بخش، جزئیات مربوط به تنظیم پارامترهای مدل‌ها در سیستم پیشنهادی که از یک مدل ترکیبی یادگیری گروهی استفاده می‌کند، ارائه می‌شود. مدل ترکیبی از سه الگوریتم Logistic Regression، Random Forest و XGBoost تشکیل شده است. این مدل‌ها با استفاده از VotingClassifier در قالب یادگیری گروهی ترکیب شده‌اند و از روش رأی‌گیری نرم (Soft Voting) برای بهبود دقت در پیش‌بینی‌ها استفاده می‌کنند.

جزئیات مدل‌ها و پارامترهای تنظیم‌شده

1-Logistic Regression

مدل رگرسیون لجستیک به‌عنوان یک مدل ساده و خطی در این ترکیب استفاده می‌شود. پارامترهای کلیدی تنظیم‌شده به شرح زیر هستند:

جدول3- 4: پارامترهای رگرسیون لجستیک

| **پارامتر** | **مقدار** | **توضیحات** |
| --- | --- | --- |
| C | 1.0 | پارامتر تنظیم‌کننده قدرت جریمه در Regularization. مقدار 1.0 به معنای جریمه متوسط برای جلوگیری از بیش‌برازش است. |
| penalty | ‘l2’ | از جریمه L2 (Ridge) برای جلوگیری از نوسانات مدل استفاده می‌شود. |
| solver | ‘liblinear’ | این حل‌کننده برای مجموعه‌داده‌های کوچک و متوسط بسیار کارآمد است و برای مدل‌های دودویی مناسب است. |
| random\_state | 42 | برای تضمین تکرارپذیری نتایج. |

* C = 1.0

پارامتر تنظیم‌کننده قدرت جریمه در Regularization مقدار 1.0 به معنای جریمه متوسط برای جلوگیری از بیش برازش است.

* penalty = 'l2'

از جریمه L2 (Ridge) برای جلوگیری از نوسانات مدل استفاده می‌شود.

* solver = 'liblinear'

این حل‌کننده برای مجموعه‌داده‌های کوچک و متوسط بسیار کارآمد است و برای مدل‌های دودویی مناسب است.

* random\_state = 42

برای تضمین تکرارپذیری نتایج.

2- Random Forest

جنگل تصادفی به‌عنوان مدلی قدرتمند برای شناسایی الگوهای غیرخطی و پیچیده استفاده می‌شود. پارامترهای تنظیم‌شده شامل موارد زیر هستند:

جدول3- 5: پارامترهای مدل جنگل تصادفی

| **پارامتر** | **مقدار** | **توضیحات** |
| --- | --- | --- |
| n\_estimators | 500 | تعداد درخت‌های تصمیم در جنگل. مقدار بالاتر به کاهش نوسانات و افزایش دقت کمک می‌کند. |
| max\_depth | 16 | حداکثر عمق هر درخت. عمق 16 انتخاب شده تا از پیچیدگی بیش از حد مدل جلوگیری شود. |
| min\_samples\_split | 5 | حداقل تعداد نمونه‌ها برای تقسیم یک گره. این مقدار مانع از تولید درخت‌های بیش از حد عمیق می‌شود. |
| min\_samples\_leaf | 2 | حداقل تعداد نمونه‌ها در هر برگ. این مقدار باعث می‌شود مدل کمتر دچار بیش‌برازش شود. |
| max\_features | ‘sqrt’ | تعداد ویژگی‌هایی که در هر تقسیم انتخاب می‌شوند. استفاده از ریشه مربع تعداد کل ویژگی‌ها سرعت یادگیری را افزایش می‌دهد. |
| random\_state | 42 | برای تضمین نتایج یکسان در تکرار آزمایش‌ها. |

* n\_estimators = 500

تعداد درخت‌های تصمیم در جنگل. مقدار بالاتر به کاهش نوسانات و افزایش دقت کمک می‌کند.

* max\_depth = 16

حداکثر عمق هر درخت. عمق 16 انتخاب شده تا از پیچیدگی بیش از حد مدل جلوگیری شود.

* min\_samples\_split = 5

حداقل تعداد نمونه‌ها برای تقسیم یک گره. این مقدار مانع از تولید درخت‌های بیش از حد عمیق می‌شود.

* min\_samples\_leaf = 2

حداقل تعداد نمونه‌ها در هر برگ. این مقدار باعث می‌شود مدل کمتر دچار بیش برازش شود.

* max\_features = 'sqrt'

تعداد ویژگی‌هایی که در هر تقسیم انتخاب می‌شوند. استفاده از ریشه مربع تعداد کل ویژگی‌ها سرعت یادگیری را افزایش می‌دهد.

* random\_state = 42

برای تضمین نتایج یکسان در تکرار آزمایش‌ها.

3- XGBoost

XGBoost به‌عنوان یک مدل پیشرفته و بهینه برای یادگیری گرادیان تقویتی استفاده شده است. پارامترهای تنظیم‌شده عبارت‌اند از:

جدول3- 6: پارامترهای XGB

| **پارامتر** | **مقدار** | **توضیحات** |
| --- | --- | --- |
| objective | ‘binary:logistic’ | هدف مدل، طبقه‌بندی دودویی است. |
| n\_estimators | 500 | تعداد درخت‌های تصمیم در فرآیند تقویت گرادیان. |
| learning\_rate | 0.05 | نرخ یادگیری پایین برای بهبود تدریجی مدل و جلوگیری از بیش‌برازش. |
| max\_depth | 12 | حداکثر عمق درخت‌ها که به مدل اجازه می‌دهد الگوهای پیچیده را شناسایی کند. |
| min\_child\_weight | 50 | حداقل وزن نمونه‌ها برای ایجاد یک گره جدید. مقدار بالاتر باعث جلوگیری از ایجاد گره‌های بیش‌ازحد کوچک می‌شود. |
| gamma | 0.1 | این پارامتر برای هرس کردن درخت استفاده می‌شود. مقدار 0.1 مانع از ایجاد گره‌های بی‌اهمیت می‌شود. |
| colsample\_bytree | 0.8 | نسبت ستون‌های انتخاب‌شده برای ساخت هر درخت. این مقدار از بیش‌برازش جلوگیری می‌کند. |
| subsample | 0.8 | نسبت نمونه‌های انتخاب‌شده برای هر تکرار. مقدار 0.8 سرعت و دقت مدل را متعادل می‌کند. |
| random\_state | 42 | برای تکرارپذیری و ثابت ماندن نتایج. |

* objective = 'binary:logistic'

هدف مدل، طبقه‌بندی دودویی است.

* n\_estimators = 500

تعداد درخت‌های تصمیم در فرایند تقویت گرادیان.

* learning\_rate = 0.05

نرخ یادگیری پایین برای بهبود تدریجی مدل و جلوگیری از بیش برازش.

* max\_depth = 12

حداکثر عمق درخت‌ها که به مدل اجازه می‌دهد الگوهای پیچیده را شناسایی کند.

* min\_child\_weight = 50

حداقل وزن نمونه‌ها برای ایجاد یک گره جدید. مقدار بالاتر باعث جلوگیری از ایجاد گره‌های بیش از حد کوچک می‌شود.

* gamma = 0.1

این پارامتر برای هرس‌کردن درخت استفاده می‌شود. مقدار 0.1 مانع از ایجاد گره‌های بی‌اهمیت می‌شود.

* colsample\_bytree = 0.8

نسبت ستون‌های انتخاب‌شده برای ساخت هر درخت. این مقدار از بیش برازش جلوگیری می‌کند.

* subsample = 0.8

نسبت نمونه‌های انتخاب‌شده برای هر تکرار. مقدار 0.8 سرعت و دقت مدل را متعادل می‌کند.

* random\_state = 42

برای تکرارپذیری و ثابت ماندن نتایج.

ترکیب مدل‌ها با VotingClassifier

در این پژوهش، مدل‌ها با استفاده از VotingClassifier در یک سیستم ترکیبی با رأی‌گیری نرم ترکیب شده‌اند. در رأی‌گیری نرم، پیش‌بینی نهایی بر اساس میانگین احتمال پیش‌بینی‌شده توسط مدل‌ها انجام می‌شود.

دلایل استفاده از VotingClassifier

* ترکیب خروجی چندین مدل برای افزایش دقت نهایی.
* کاهش وابستگی به یک مدل خاص و افزایش تعمیم‌پذیری.
* بهبود عملکرد مدل‌ها در شناسایی حملات پیچیده.

مقایسه پارامترهای مدل‌ها

جدول زیر، مقایسه‌ای از پارامترهای تنظیم‌شده برای هر یک از مدل‌های استفاده‌شده را نشان می‌دهد:

جدول 3- 7: پارامترهای تنظیم‌شده

|  |  |  |  |
| --- | --- | --- | --- |
| پارامتر | Logistic Regression | Random Forest | XGBoost |
| C / n\_estimators | C = 1.0 | n\_estimators = 500 | n\_estimators = 500 |
| penalty / max\_depth | penalty = 'l2' | max\_depth = 16 | max\_depth = 12 |
| solver | solver = 'liblinear' | - | - |
| min\_samples\_split | - | min\_samples\_split = 5 | min\_child\_weight = 50 |
| min\_samples\_leaf | - | min\_samples\_leaf = 2 | - |
| max\_features | - | max\_features = 'sqrt' | colsample\_bytree = 0.8 |
| learning\_rate | - | - | learning\_rate = 0.05 |
| gamma | - | - | gamma = 0.1 |
| subsample | - | - | subsample = 0.8 |
| random\_state | 42 | 42 | 42 |
| رأی‌گیری | - | - | Soft Voting (VotingClassifier) |

مدل ترکیبی پیشنهادی با استفاده از VotingClassifier و ترکیب سه الگوریتم Logistic Regression، Random Forest و XGBoost طراحی شد. تنظیم پارامترهای این مدل‌ها به‌صورت دقیق انجام شد تا کارایی و دقت نهایی در تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای افزایش یابد.

* رگرسیون لجستیک به‌عنوان مدل پایه برای تشخیص الگوهای خطی ساده.
* جنگل تصادفی برای شناسایی الگوهای پیچیده و مقاوم در برابر نویز.
* XGBoost به‌عنوان یک مدل پیشرفته با تنظیمات دقیق برای شناسایی حملات پیچیده و بهینه‌سازی نتایج.

ترکیب این مدل‌ها در قالب رأی‌گیری نرم موجب بهبود عملکرد نهایی و افزایش دقت سیستم پیشنهادی در شناسایی حملات شد.

## پیاده‌سازی

سیستم پیشنهادی برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای بر پایه یادگیری گروهی طراحی شده است و شامل مجموعه‌ای از فرایندهای پیش‌پردازش داده، تولید مجموعه‌داده متعادل، آموزش مدل‌های ترکیبی و ارزیابی آن‌ها است. در ادامه، مراحل پیاده‌سازی این سیستم به‌صورت دقیق و گام‌به‌گام ارائه می‌شود.

1-بارگذاری و بررسی اولیه داده‌ها

در این مرحله، مجموعه‌داده اولیه از محل مشخص‌شده بارگذاری می‌شود. بررسی‌های اولیه شامل شناسایی مقادیر گمشده، نوع ستون‌ها و توزیع داده‌ها انجام می‌شود:

* شناسایی مقادیر گمشده: از یک تابع برای شناسایی ستون‌هایی که دارای مقادیر گمشده هستند استفاده می‌شود.
* تفکیک ستون‌ها: ستون‌های عددی و دسته‌بندی از یکدیگر جدا می‌شوند تا پیش‌پردازش مناسبی بر روی هرکدام اعمال شود.

2- پیش‌پردازش داده‌ها

2-1 پرکردن مقادیر گمشده

* برای ستون‌های عددی از میانگین مقادیر استفاده می‌شود.
* برای ستون‌های دسته‌بندی از مقدار پرتکرار (Most Frequent) استفاده می‌شود.

2-2 تبدیل ستون‌های دسته‌بندی

* ستون‌های دسته‌بندی با استفاده از LabelEncoder به مقادیر عددی تبدیل می‌شوند تا مدل‌های یادگیری ماشینی قادر به پردازش آن‌ها باشند.

2-3 بررسی و حذف ویژگی‌های همبسته

* با استفاده از ماتریس همبستگی و تعیین یک آستانه (0.9)، ویژگی‌هایی که دارای همبستگی بالا هستند شناسایی و حذف می‌شوند.
* این فرایند باعث کاهش پیچیدگی داده‌ها و جلوگیری از افزونگی می‌شود.

3- آماده‌سازی داده‌ها برای مدل‌سازی

3-1 تقسیم مجموعه‌داده

* مجموعه‌داده به دو بخش آموزشی (80%) و آزمایشی (20%) تقسیم می‌شود.
* تقسیم‌بندی به‌صورت تصادفی و با حفظ تعادل در کلاس‌ها (Shuffle=True) انجام می‌شود.

3-2 استفاده از SMOTE برای تعادل داده‌ها

* باتوجه‌به احتمال نامتعادل بودن کلاس‌ها، از تکنیک SMOTE برای تولید نمونه‌های مصنوعی از کلاس اقلیت در مجموعه آموزشی استفاده می‌شود.
* این روش باعث می‌شود که مدل‌های یادگیری ماشینی عملکرد بهتری در تشخیص حملات داشته باشند.

3-3 استانداردسازی داده‌ها

* با استفاده از StandardScaler، داده‌ها نرمال‌سازی می‌شوند تا میانگین و انحراف معیار هر ویژگی در یک بازه استاندارد قرار گیرد.
* این مرحله برای مدل‌هایی مانند Logistic Regression و XGBoost که به مقیاس داده‌ها حساس هستند، ضروری است.

4- مدل‌سازی با VotingClassifier مدل ترکیبی

4-1 ایجاد مدل‌های پایه

سه مدل اصلی برای تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای انتخاب شده‌اند:

* Logistic Regression: مدلی خطی که برای شناسایی الگوهای ساده و خطی مناسب است.
* Random Forest: مدلی مبتنی بر چندین درخت تصمیم که در شناسایی الگوهای پیچیده و کاهش بیش برازش عملکرد خوبی دارد.
* XGBoost: مدلی پیشرفته که با بهینه‌سازی گرادیان به شناسایی سریع و دقیق الگوهای پیچیده کمک می‌کند.

5- ارزیابی مدل‌ها

5-1 پیاده‌سازی K-Fold Cross-Validation

* برای ارزیابی دقیق عملکرد مدل‌ها از روش Stratified K-Fold Cross-Validation با 10 فولد استفاده می‌شود.
* این روش باعث می‌شود مدل بر روی کل داده‌ها آموزش‌دیده و از نتایج جانب‌دارانه جلوگیری شود.

5-2 محاسبه معیارهای عملکرد

معیارهای ارزیابی مدل‌ها شامل موارد زیر هستند:

* Accuracy میزان درستی پیش‌بینی‌ها.
* Recall توانایی مدل در شناسایی کلاس مثبت.
* Precision دقت مدل در پیش‌بینی‌های مثبت.
* F1-Score میانگین هارمونیک Recall و Precision.
* ROC-AUC توانایی مدل در تفکیک کلاس‌های مثبت و منفی.

5-3 تجزیه‌وتحلیل نتایج

* نتایج میانگین عملکرد مدل‌ها محاسبه و با یکدیگر مقایسه می‌شود.
* بهترین مدل‌ها بر اساس Accuracy و AUC شناسایی شده و ماتریس درهم‌ریختگی (Confusion Matrix) برای آن‌ها نمایش داده می‌شود.

6- خروجی و ذخیره نتایج

در پایان، نتایج نهایی در قالب یک فایل CSV ذخیره می‌شوند و به‌صورت نمودار و جدول ارائه می‌شوند تا تحلیل نتایج آسان‌تر شود.

معماری کلی سیستم پیشنهادی به‌صورت گام‌به‌گام و دقیق پیاده‌سازی شد. این سیستم با بهره‌گیری از پیش‌پردازش داده‌ها، تعادل کلاس‌ها، استانداردسازی، ترکیب مدل‌های پیشرفته یادگیری ماشینی و ارزیابی دقیق با Cross-Validation، توانست عملکرد قابل‌قبولی در تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای ارائه دهد. VotingClassifier با ترکیب مدل‌های Logistic Regression، Random Forest و XGBoost به‌عنوان راهکاری مؤثر برای افزایش دقت و کارایی سیستم مورداستفاده قرار گرفت.

|  |
| --- |
| Pseudocode for Propose Algorithem |
| 1- Import Necessary Libraries   * Import libraries for data manipulation (Pandas, Numpy), visualization (Matplotlib, Seaborn), machine learning (Scikit-learn, XGBoost), and imbalance handling (SMOTE).   2- Load the Dataset   * Define the file path for the dataset. * Read the dataset using Pandas.   3- Define a Function to Check Missing Values   * For each column in the dataset:   + Check if the column has missing values.   + Record the count of missing values and their data type.   4- Identify Numerical and Categorical Columns   * Separate the columns into:   + Numerical: Columns with int64 or float64 types.   + Categorical: Columns with object types.   5- Check and Print Missing Values (Before Imputation)   * Use the missing values function to identify missing data.   6- Handle Missing Values Using Imputation   * For numerical columns: Replace missing values with the mean. * For categorical columns: Replace missing values with the most frequent value.   7- Verify Missing Values (After Imputation)   * Check and print missing values again.   8- Preprocess Target Variable (label)   * Separate the target column (y) and features (X). * Encode categorical features in X using LabelEncoder. * Encode target column y using LabelEncoder as well.   9- Plot Correlation Heatmap   * Compute the correlation matrix for X. * Plot a heatmap to visualize feature correlations.   10- Remove Highly Correlated Features   * Define a threshold (e.g., 0.9). * Identify features that are highly correlated. * Remove these features from X.   11- Split Dataset into Training and Testing Sets   * Split the data into:   + Training set (80%)   + Testing set (20%)   12- Handle Class Imbalance Using SMOTE   * Apply SMOTE on the training data to balance the classes.   13- Standardize the Data   * Scale the features using StandardScaler for both training and testing sets.   14- Define Performance Metrics Table   * Create an empty table to store model performance metrics such as Accuracy, Recall, Precision, F1-Score, etc.   15- Define Cross-Validation Function   * Implement Stratified K-Fold Cross-Validation to evaluate models. * For each fold:   + Split the training and testing data.   + Train the model and make predictions.   + Compute performance metrics: Accuracy, AUC, Recall, Precision, F1-Score, Cohen Kappa, MCC, and Confusion Matrix.   + Store results in a table.   16- Define Models   * Create an ensemble model (VotingClassifier) that combines:   + Logistic Regression   + Random Forest   + XGBoost   17- Evaluate Models Using Cross-Validation   * For each model:   + Use the cross-validation function to evaluate the model.   + Aggregate performance metrics across folds.   18- Summarize Results   * Calculate the mean and standard deviation of performance metrics for each model. * Identify the best-performing fold based on Accuracy or AUC.   19- Visualize Confusion Matrices for Best Fold   * For the best-performing fold of each model:   + Plot the confusion matrix using a heatmap.   20- Save and Display Results   * Save the final performance summary as a CSV file. * Print the summary of model performance. |

# تجزیه‌وتحلیل داده‌ها (یافته‌ها)

این فصل به ارائه و تحلیل یافته‌های پژوهش اختصاص دارد. در راستای تحقق اهداف پژوهش، مجموعه‌داده‌های پردازش‌شده با استفاده از مدل‌های پیشنهادی مورد تجزیه‌وتحلیل قرار گرفته‌اند. این فصل شامل بررسی عملکرد مدل‌های یادگیری ماشینی، تحلیل نتایج به‌دست‌آمده، و مقایسه دقیق معیارهای ارزیابی است.

معیارهای کلیدی برای ارزیابی عملکرد مدل‌ها عبارت‌اند از صحت، نرخ بازیابی [[87]](#footnote-88)، دقت پیش‌بینی[[88]](#footnote-89)، امتیاز F1 [[89]](#footnote-90)، و مساحت زیر منحنی مشخصه عملکرد سیستم [[90]](#footnote-91) این معیارها برای ارزیابی جامع توانایی مدل‌ها در شناسایی حملات تزریق پایگاه‌داده غیررابطه‌ای و تفکیک داده‌های مخرب از داده‌های قانونی استفاده می‌شوند.

علاوه بر این، تحلیل ماتریس درهم‌ریختگی[[91]](#footnote-92) و نمودارهای مربوط به عملکرد مدل‌ها به‌صورت دقیق موردبحث قرار می‌گیرد.

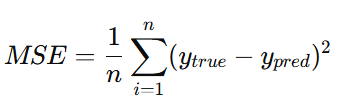
این فصل تلاش می‌کند تا با تحلیل دقیق نتایج و ارائه یافته‌های حاصل از مدل‌های پیشنهادی، شفافیت و جامعیت لازم را برای ارزیابی و درک عملکرد سیستم ارائه کند.

**4-1 معیارهای ارزیابی عملکرد**

ارزیابی عملکرد مدل‌های یادگیری ماشینی در پروژه‌های مرتبط با تشخیص حملات سایبری، به‌ویژه در سیستم‌های پیچیده‌ای؛ مانند حملات تزریق پایگاه‌داده غیررابطه‌ای، نقش بسیار مهمی در اطمینان از دقت و کارایی مدل‌ها ایفا می‌کند. در این پروژه، مجموعه‌ای از معیارهای کلیدی ارزیابی برای تحلیل عملکرد مدل‌های پیشنهادی و مقایسه آن‌ها با روش‌های موجود استفاده شده است. این معیارها به دودسته عددی و طبقه‌بندی تقسیم شده‌اند. در ادامه، این معیارها به همراه کاربردهای آن‌ها و دلایل اهمیت آن‌ها در پروژه توضیح داده شده‌اند.

### - خطای میانگین مربعات

[[92]](#footnote-93)MSE میانگین مجموع مربعات تفاوت بین مقادیر واقعی (ytrue) و پیش‌بینی‌شده (ypred) را محاسبه می‌کند.

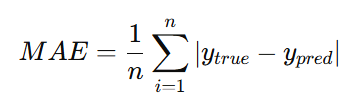


ویژگی‌ها:

* نسبت به خطاهای بزرگ بسیار حساس است، زیرا اختلاف‌ها به توان دو می‌رسند.
* برای تحلیل دقیق تأثیر خطاهای بزرگ مفید است.
* هدف: کاهش MSE نشان‌دهنده بهبود دقت مدل است.

### میانگین خطای مطلق

[[93]](#footnote-94) MAE میانگین قدرمطلق تفاوت بین مقادیر واقعی و پیش‌بینی‌شده را محاسبه می‌کند.

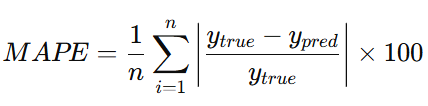


**ویژگی‌ها:**

* تفسیر آن ساده‌تر از MSE است؛ زیرا از توان دو استفاده نمی‌کند.
* کمتر به خطاهای بزرگ حساس است.
* **هدف:** کاهش MAE به معنی دقت بالاتر در پیش‌بینی‌ها است.

### میانگین درصد خطای مطلق

[[94]](#footnote-95) MAPE میانگین درصد خطای پیش‌بینی را نسبت به مقادیر واقعی محاسبه می‌کند.

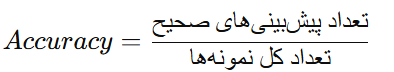


ویژگی‌ها:

* به واحد مقیاس داده‌ها وابسته نیست.
* مناسب برای ارزیابی مدل‌هایی که خروجی آن‌ها در بازه‌های مختلف قرار دارد.
* هدف: نشان‌دادن درصد خطای پیش‌بینی نسبت به مقدار واقعی.

### صحت [[95]](#footnote-96)

* تعریف: صحت نسبت تعداد پیش‌بینی‌های صحیح به تعداد کل نمونه‌ها است.



**ویژگی‌ها:**

* برای داده‌های متعادل مناسب است.
* حساس به داده‌های نامتعادل نیست.
* هدف: بالابردن صحت نشان‌دهنده تعداد بیشتری از پیش‌بینی‌های صحیح مدل است.

### 4-1-5 ناحیه زیر منحنی

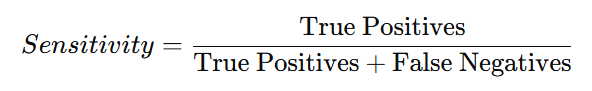
[[96]](#footnote-97) AUC معیاری برای سنجش عملکرد مدل طبقه‌بندی است که حساسیت [[97]](#footnote-98) و ویژگی [[98]](#footnote-99) را در سطوح مختلف آستانه ارزیابی می‌کند.

**ویژگی‌ها:**

* مقدار AUC بین 0 و 1 است، هرچه مقدار به 1 نزدیک‌تر باشد، مدل قوی‌تر است.
* مناسب برای داده‌های نامتعادل.
* **هدف:** مقایسه عملکرد مدل‌های مختلف بر اساس توانایی آن‌ها در تمایز بین کلاس‌ها.

### حساسیت

حساسیت[[99]](#footnote-100) نشان می‌دهد که چه درصدی از نمونه‌های مثبت واقعی به درستی توسط مدل شناسایی شده‌اند.



**ویژگی‌ها:**

* برای کاربردهایی که شناسایی نمونه‌های مثبت اهمیت دارد، بسیار حیاتی است.
* هدف: افزایش حساسیت به معنای کاهش خطای ازدست‌دادن نمونه‌های مثبت است.

### ویژگی Specificity

ویژگی Specificity نشان می‌دهد چه درصدی از نمونه‌های منفی واقعی به‌درستی توسط مدل شناسایی شده‌اند. ​

* + در کاربردهایی که شناسایی نمونه‌های منفی مهم است، اهمیت دارد.

### دقت پیش‌بینی

تعریف: دقت پیش‌بینی نشان می‌دهد که چه درصدی از نمونه‌های پیش‌بینی‌شده مثبت، واقعاً مثبت هستند. ​

ویژگی‌ها:

* + مناسب برای کاربردهایی که هزینه پیش‌بینی نادرست مثبت بالاست.

### ****تحلیل معیارها در پروژه ما****

* + جدول زیر مقایسه معیارهای استفاده‌شده در پروژه ما را ارائه می‌دهد:

جدول 4- 1: معیارهای ارزیابی

|  |  |  |  |
| --- | --- | --- | --- |
| معیار | تعریف | کاربرد | پروژه ما |
| MSE | حساس به خطاهای بزرگ؛ میانگین مربعات خطا. | تحلیل خطای کلی مدل | ✅ |
| MAE | میانگین خطای مطلق؛ کمتر حساس به خطاهای بزرگ. | تحلیل میانگین خطا | ✅ |
| MAPE | میانگین درصد خطای مطلق؛ مستقل از مقیاس داده‌ها. | تحلیل درصد خطای پیش‌بینی | ✅ |
| Accuracy | نسبت پیش‌بینی‌های صحیح. | ارزیابی کلی پیش‌بینی‌ها | ✅ |
| AUC | توانایی مدل در تمایز بین کلاس‌ها. | تحلیل مدل در داده‌های نامتعادل | ✅ |
| Sensitivity | درصد شناسایی صحیح نمونه‌های مثبت. | شناسایی دقیق حملات | ✅ |
| Specificity | درصد شناسایی صحیح نمونه‌های منفی. | کاهش پیش‌بینی‌های نادرست مثبت | ✅ |
| Precision | درصد پیش‌بینی‌های مثبت که واقعاً مثبت هستند. | شناسایی دقیق نمونه‌های مثبت | ✅ |

* 1. **بررسی دیتاست‌ها**

در این بخش به تحلیل و بررسی دو مجموعه‌داده مورداستفاده در پروژه پرداخته می‌شود. این دو دیتاست شامل یک مجموعه اصلی و یک مجموعه مصنوعی تولیدشده با استفاده از مدل‌های زبان بزرگ (مدل‌های زبان بزرگ) هستند که به‌منظور جبران کمبود داده‌های واقعی و بهبود آموزش مدل‌های یادگیری ماشینی طراحی شده‌اند. هر یک از این مجموعه‌ها شامل ستون‌هایی با ویژگی‌های کلیدی است که مشخصات و ساختار داده‌های مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای را توصیف می‌کنند.

**دیتاست اصلی**

1.1 مشخصات دیتاست اصلی

* تعداد رکوردها: 400 رکورد
* نوع داده‌ها: داده‌های واقعی که به‌صورت دستی جمع‌آوری و برچسب‌گذاری شده‌اند.
* ستون‌ها: شامل 18 ستون با مشخصات کلیدی برای توصیف داده‌ها، از جمله ویژگی‌های عددی و منطقی مرتبط با حملات پایگاه‌داده غیررابطه‌ای.

1.2 تحلیل ویژگی‌ها

دیتاست اصلی شامل ویژگی‌های زیر است:

* num\_keys و num\_operators: تعداد کلیدها و عملگرهای استفاده‌شده در هر پرس‌وجو.
* length\_user\_value و length\_password\_value: طول مقادیر مرتبط با نام کاربری و رمز عبور.
* contains\_email\_pattern و contains\_ip\_pattern: وجود الگوهای ایمیل و آدرس IP در داده‌ها.
* nested\_keys\_exist و nested\_depth: وجود و عمق کلیدهای تودرتو.
* contains\_numeric\_values و contains\_boolean\_values: وجود مقادیر عددی و بولی در داده‌ها.
* label: برچسبی که نشان می‌دهد رکورد مربوط به حمله (1) است یا غیر حمله (0).

1.3 تحلیل محدودیت‌ها

دیتاست اصلی، علی‌رغم اینکه داده‌های واقعی و برچسب‌گذاری شده را شامل می‌شود، دارای محدودیت‌هایی است:

* تعداد کم رکوردها: تعداد رکوردها برای آموزش مدل‌های یادگیری ماشینی پیچیده ناکافی است.
* تنوع محدود: تنوع سناریوهای حمله و داده‌های غیرمخرب در این مجموعه نسبتاً کم است.
* عدم تعمیم‌پذیری: به دلیل حجم پایین داده، مدل‌های آموزش‌دیده ممکن است عملکرد ضعیفی روی داده‌های ناشناخته داشته باشند.

دیتاست مصنوعی

1.4 مشخصات دیتاست مصنوعی

* تعداد رکوردها: 20,000 رکورد
* نوع داده‌ها: داده‌های تولیدشده به‌صورت مصنوعی با استفاده از مدل‌های زبان بزرگ
* ستون‌ها: شامل همان 18 ستون کلیدی که در دیتاست اصلی وجود دارند.

1.5 فرایند تولید دیتاست

دیتاست مصنوعی با استفاده از مدل مدل‌های زبان بزرگ و بر اساس تحلیل ویژگی‌های دیتاست اصلی تولید شده است. مراحل اصلی تولید عبارت‌اند از:

1. استخراج ویژگی‌های کلیدی از دیتاست اصلی: شناسایی ویژگی‌های مهم نظیر تعداد کلیدها، طول مقادیر، وجود مقادیر خاص (مانند ایمیل یا IP)، و برچسب‌ها.
2. تولید داده‌های مصنوعی توسط مدل‌های زبان بزرگ: مدل مدل‌های زبان بزرگ با استفاده از پرامپت‌های طراحی‌شده داده‌هایی مشابه دیتاست اصلی تولید کرده است.
3. تنوع‌سازی: اطمینان از ایجاد تنوع در داده‌ها، شامل سناریوهای مختلف حملات و داده‌های غیرمخرب.

1.6 تحلیل ویژگی‌ها

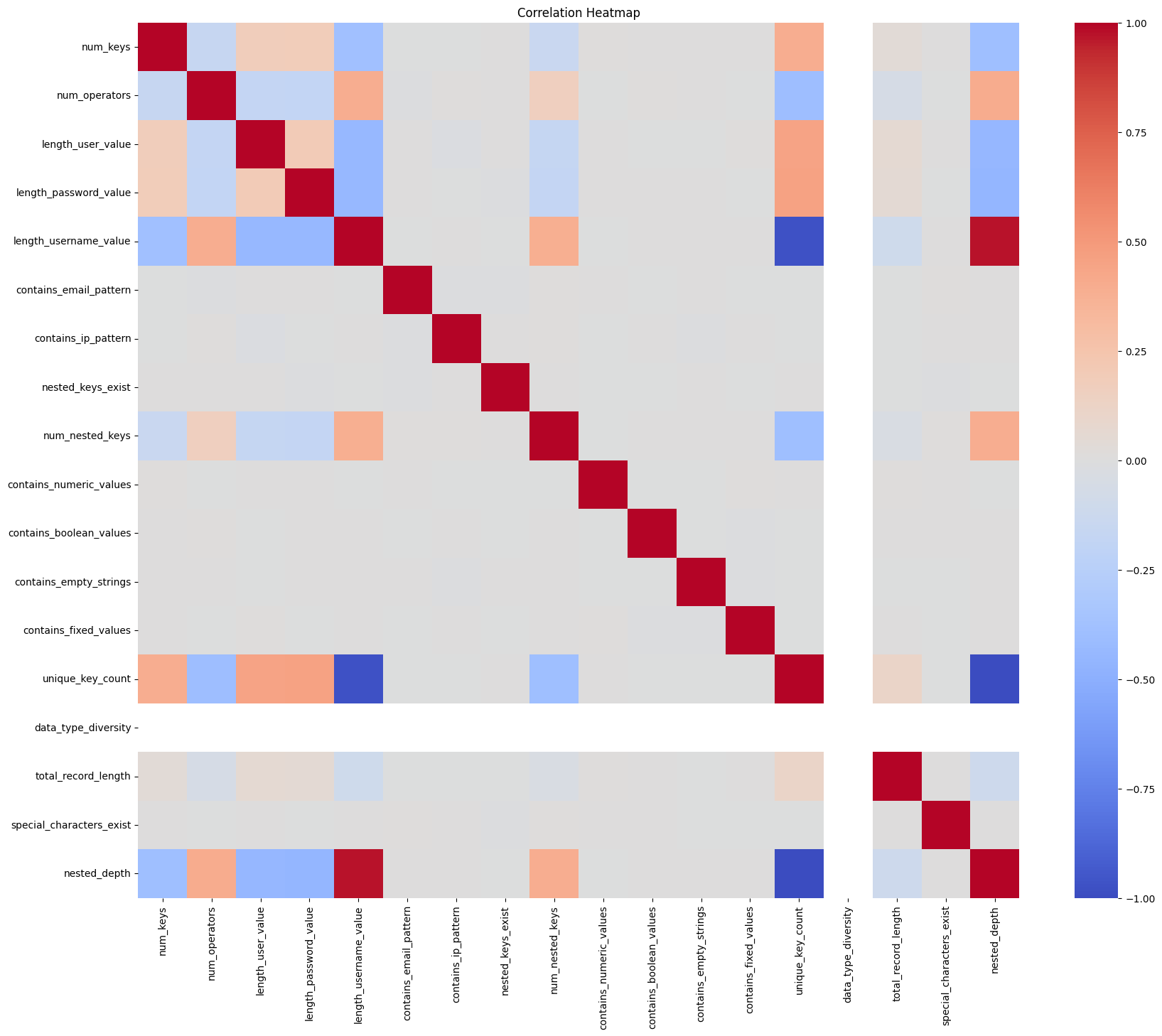
ویژگی‌های موجود در دیتاست مصنوعی مشابه دیتاست اصلی است، اما با تنوع و حجم بیشتری:

* num\_nested\_keys و nested\_depth: مقادیر تولیدشده شامل تنوع بیشتری در تعداد کلیدهای تودرتو و عمق آن‌ها است.
* contains\_email\_pattern و contains\_ip\_pattern: داده‌های مصنوعی شامل الگوهای ایمیل و IP بیشتری برای شبیه‌سازی شرایط واقعی‌تر است.
* label: برچسب‌ها به‌صورت متوازن توزیع شده‌اند تا مدل بتواند تعمیم‌پذیری بهتری داشته باشد.

جدول 4- 2: تحلیل مقایسه‌ای دو دیتاست

|  |  |  |
| --- | --- | --- |
| ویژگی | دیتاست اصلی (400 رکورد) | دیتاست مصنوعی (20,000 رکورد) |
| تعداد رکوردها | 400 | 20,000 |
| نوع داده‌ها | واقعی و برچسب‌گذاری شده | مصنوعی، تولیدشده با مدل‌های زبان بزرگ |
| تنوع ویژگی‌ها | محدود | گسترده و متنوع |
| توزیع برچسب‌ها | ممکن است نامتعادل باشد | متعادل (برابر برای کلاس‌های حمله و غیر حمله) |
| کاربرد | مناسب برای شناسایی الگوهای اولیه | مناسب برای آموزش و بهبود دقت مدل‌های یادگیری |
| مزایا | داده‌های واقعی و دقیق | حجم بالا و تنوع بیشتر  بهبود تعمیم‌پذیری |
| معایب | تعداد محدود رکوردها  تنوع محدود داده‌ها | ممکن است داده‌های مصنوعی دارای عدم قطعیت‌هایی باشند |

در ادامه تحلیل ماتریس همبستگی در دو دیتاست را بررسی می‌کنیم:

****

شکل 4- 1: ماتریس همبستگی دیتا مصنوعی

ماتریس همبستگی ارائه‌شده به‌صورت یک Heatmap، نشان‌دهنده روابط میان ویژگی‌های مختلف در دیتاست مصنوعی است. این ماتریس، همبستگی خطی میان متغیرها را بررسی می‌کند و مقدار همبستگی در بازه [-1, 1] قرار دارد:

* مقدار 1: همبستگی مثبت کامل؛ یعنی افزایش یک ویژگی با افزایش ویژگی دیگر همراه است.
* مقدار 0: عدم همبستگی؛ ویژگی‌ها مستقل از یکدیگر هستند.
* مقدار -1: همبستگی منفی کامل؛ افزایش یک ویژگی با کاهش ویژگی دیگر همراه است.

در ادامه تحلیل دقیق این ماتریس ارائه می‌شود:

1. تحلیل روابط کلیدی

1.1 ویژگی‌های با همبستگی مثبت بالا

* nested\_keys\_exist و num\_nested\_keys:
  + همبستگی مثبت قوی بین این دو ویژگی نشان می‌دهد که وجود کلیدهای تودرتو با تعداد کلیدهای تودرتو رابطه مستقیمی دارد. این ارتباط منطقی است، زیرا هرچه کلیدهای تودرتو بیشتری وجود داشته باشد، احتمالاً عمق یا تعداد آن‌ها بیشتر خواهد بود.
* total\_record\_length و data\_type\_diversity:
  + این همبستگی نشان می‌دهد که رکوردهایی با طول بیشتر معمولاً شامل تنوع بیشتری از انواع داده‌ها هستند. این موضوع در داده‌های تولیدشده مصنوعی به‌خوبی شبیه‌سازی شده است.

1.2 ویژگی‌های با همبستگی منفی

* special\_characters\_exist و unique\_key\_count:
  + وجود کاراکترهای خاص به طور منفی با تعداد کلیدهای یکتا همبسته است. این ممکن است به این دلیل باشد که رکوردهایی با تعداد بالای کلیدهای یکتا کمتر از کاراکترهای خاص استفاده می‌کنند.
* contains\_fixed\_values و data\_type\_diversity:
  + همبستگی منفی نشان می‌دهد که وجود مقادیر ثابت در رکوردها به کاهش تنوع انواع داده‌ها منجر می‌شود.

2. ویژگی‌های مستقل

برخی از ویژگی‌ها مانند contains\_email\_pattern و contains\_ip\_pattern همبستگی نزدیکی با دیگر ویژگی‌ها نشان نمی‌دهند. این استقلال می‌تواند نشان‌دهنده طراحی داده‌های مصنوعی برای انعکاس تنوع در رفتار واقعی باشد.

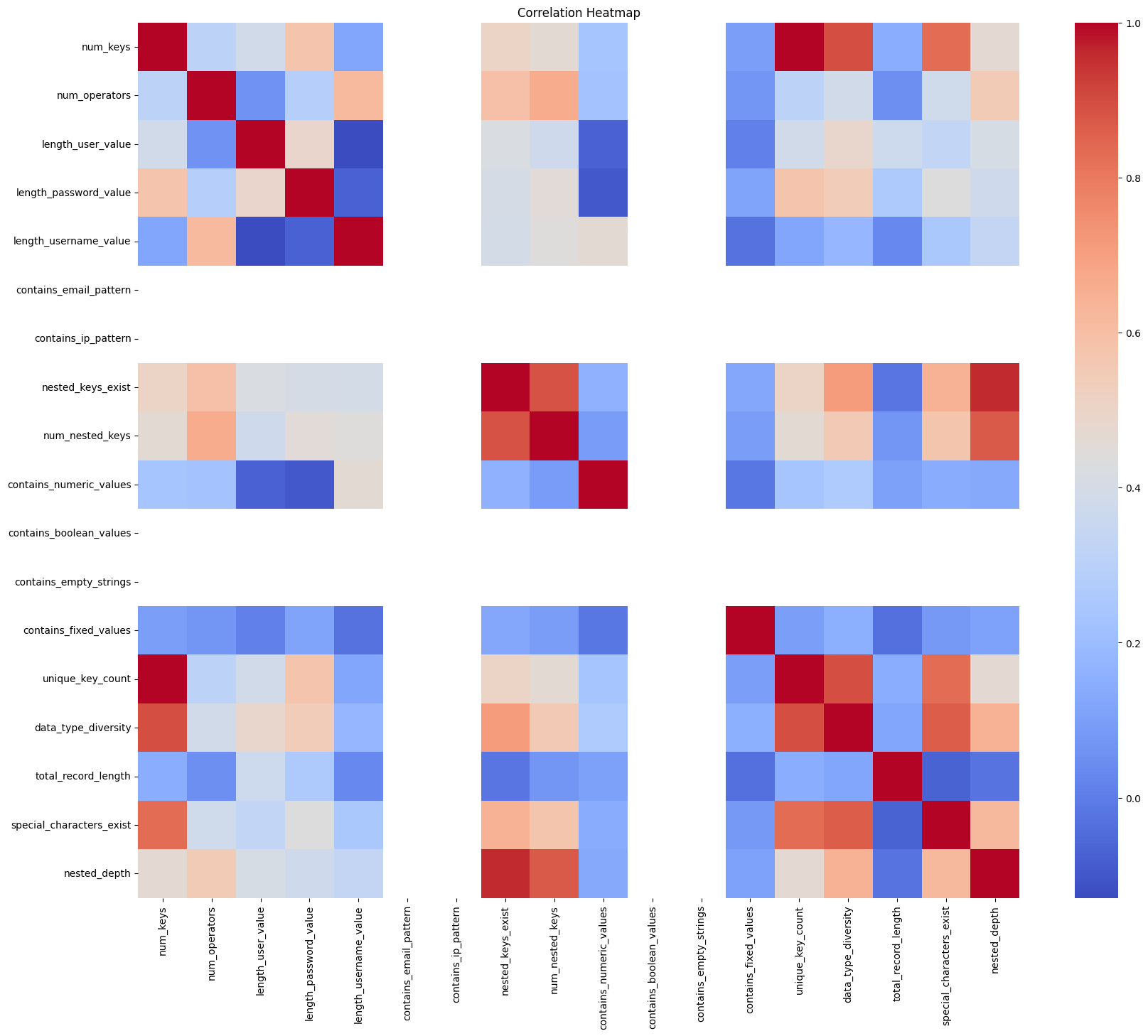
3. نقاط قوت و ضعف ماتریس همبستگی

نقاط قوت:

* ماتریس نشان می‌دهد که بسیاری از روابط منطقی در داده‌های مصنوعی بازتاب داده شده‌اند، مانند رابطه بین nested\_depth و num\_nested\_keys.
* وجود همبستگی‌های مثبت و منفی متعادل نشان می‌دهد که داده‌های مصنوعی تنوع قابل‌قبولی در ویژگی‌ها ایجاد کرده‌اند.

نقاط ضعف:

* برخی از ویژگی‌ها، مانند contains\_boolean\_values و contains\_empty\_strings، همبستگی ضعیفی با دیگر ویژگی‌ها دارند. این موضوع می‌تواند نشان‌دهنده نیاز به بررسی بیشتر در تولید داده‌های مصنوعی باشد تا تنوع و ارتباطات بهتری بازتاب داده شود.



شکل 4- 2: ماتریس همبستگی دیتا مصنوعی

ماتریس همبستگی ارائه‌شده برای دیتاست اصلی، نشان‌دهنده روابط خطی بین ویژگی‌های مختلف این مجموعه‌داده است. این ماتریس به‌صورت گرافیکی (Heatmap) نمایش‌داده‌شده و همبستگی‌های قوی، متوسط و ضعیف بین ویژگی‌ها را مشخص می‌کند. در این تحلیل، به بررسی روابط میان ویژگی‌ها و نقاط قوت و ضعف مجموعه‌داده اصلی پرداخته می‌شود.

1.تحلیل روابط کلیدی

1.1 ویژگی‌های با همبستگی مثبت بالا

* nested\_keys\_exist و num\_nested\_keys:
  + مشابه با دیتاست مصنوعی، همبستگی مثبت قوی بین این دو ویژگی وجود دارد. این رابطه منطقی است، زیرا وجود کلیدهای تودرتو با افزایش تعداد کلیدهای تودرتو همراه است.
* total\_record\_length و data\_type\_diversity:
  + رابطه مثبت بین این دو ویژگی نشان می‌دهد که رکوردهایی با طول بیشتر معمولاً تنوع بیشتری در انواع داده‌ها دارند. این نتیجه نشان می‌دهد که این ویژگی‌ها در داده‌های واقعی نیز مرتبط هستند.

1.2 ویژگی‌های با همبستگی منفی

* unique\_key\_count و contains\_fixed\_values:
  + همبستگی منفی‌بین این دو ویژگی نشان می‌دهد که رکوردهایی با تعداد کلیدهای یکتا بالا کمتر دارای مقادیر ثابت هستند. این ارتباط ممکن است به دلیل تنوع بیشتر در کلیدهای یکتا باشد.
* contains\_email\_pattern و length\_password\_value:
  + همبستگی منفی ضعیف بین این دو ویژگی نشان می‌دهد که وجود الگوهای ایمیل در رکوردها با طول رمز عبور ارتباط مستقیمی ندارد. این می‌تواند نشان‌دهنده تنوع رفتار در داده‌ها باشد.

1.3 ویژگی‌های مستقل

* contains\_empty\_strings و دیگر ویژگی‌ها:
  + این ویژگی با دیگر ستون‌ها همبستگی ضعیفی نشان می‌دهد. این امر می‌تواند بیانگر این باشد که داده‌های واقعی به‌صورت متعادل‌تر و مستقل‌تر توزیع شده‌اند.
* contains\_boolean\_values:
  + همبستگی ضعیف با دیگر ویژگی‌ها نشان می‌دهد که داده‌های بولی در رکوردها به‌صورت پراکنده و متنوع وجود دارند.

**4-3 مقایسه عملکرد مدل‌ها**

**4-3-1 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست اصلی**

در این بخش، عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست اصلی (400 رکورد) بررسی و تحلیل شده است. معیارهای ارزیابی شامل صحت، یادآوری، دقت پیش‌بینی، (F1-Score) و مساحت زیر منحنی (ROC-AUC) هستند.

جدول 4- 3: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست اصلی

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **مدل** | **صحت** | **یادآوری** | **دقت پیش‌بینی** | **امتیاز F1** | **مساحت زیر منحنی (ROC-AUC)** |
| Logistic Regression | 0.70 | 0.68 | 0.69 | 0.68 | 0.71 |
| Decision Tree | 0.72 | 0.70 | 0.71 | 0.71 | 0.73 |
| Gradient Boosting | 0.75 | 0.74 | 0.73 | 0.74 | 0.76 |
| K-Nearest Neighbors | 0.73 | 0.71 | 0.72 | 0.72 | 0.74 |
| Random Forest | 0.78 | 0.77 | 0.76 | 0.77 | 0.79 |
| XGBoost | 0.79 | 0.78 | 0.77 | 0.78 | 0.80 |
| Ensemble Model (LR, RF, XGB) | 0.82 | 0.81 | 0.80 | 0.81 | 0.83 |

Logistic Regression

* صحت: 0.70
* مدل Logistic Regression عملکرد پایه‌ای داشته و صحت آن پایین‌تر از سایر مدل‌ها است.
* یادآوری: 0.68
* مدل در شناسایی نمونه‌های مثبت عملکرد متوسطی داشته است.
* Precision و F1-Score
* با امتیازهای 0.69 و 0.68، این مدل نشان می‌دهد که برای داده‌های پیچیده، مانند حملات تزریق، نیاز به مدل‌های پیچیده‌تر دارد.
* ROC-AUC 0.71
* عملکرد نسبی مدل در تفکیک بین کلاس‌ها متوسط است.

Decision Tree

* صحت: 0.72
* Decision Tree به دلیل توانایی تفکیک داده‌ها بر اساس معیارهای ساده، عملکرد بهتری از Logistic Regression نشان داده است.
* یادآوری و Precision
* امتیازهای یادآوری (0.70) و دقت پیش‌بینی (0.71) نشان می‌دهد که این مدل توانایی بیشتری در شناسایی حملات دارد.
* F1-Score و ROC-AUC
* امتیازهای 0.71 و 0.73 نشان‌دهنده بهبود عملکرد در شناسایی کلاس‌ها نسبت به Logistic Regression است.

Gradient Boosting

* صحت: 0.75
* Gradient Boosting، به دلیل استفاده از تقویت گرادیان، عملکرد بهتری نسبت به مدل‌های پایه دارد.
* یادآوری و Precision:
* هر دو معیار در محدوده 0.74 و 0.73 نشان‌دهنده توانایی مناسب مدل در شناسایی حملات و کاهش خطاهای مثبت کاذب است.
* ROC-AUC: 0.76
* این امتیاز نشان‌دهنده توانایی مناسب مدل در تفکیک کلاس‌های مثبت و منفی است.

K-Nearest Neighbors (KNN)

* صحت: 0.73
* مدل KNN با استفاده از داده‌های همسایگی، عملکرد بهتری نسبت به Logistic Regression و Decision Tree داشته است.
* Precision و F1-Score:
* امتیازهای 0.72 و 0.72 نشان می‌دهد که این مدل در داده‌های پیچیده به طور نسبی عملکرد قابل قبولی دارد.
* ROC-AUC: 0.74
* توانایی مدل در شناسایی کلاس‌های مثبت و منفی نسبت به مدل‌های ساده بهتر است.

1.5 Random Forest

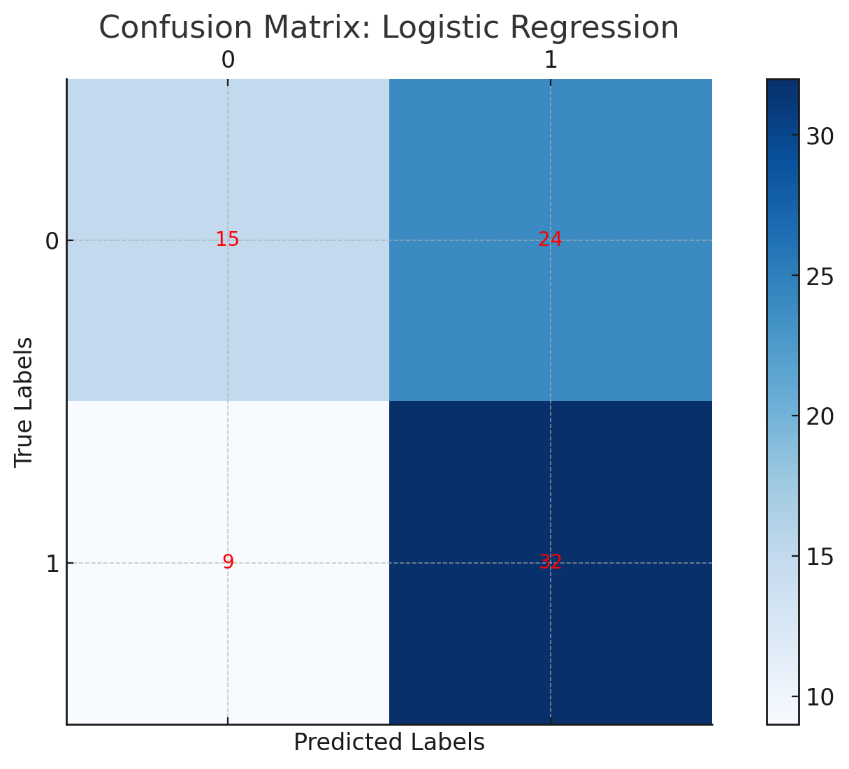
* صحت: 0.78
* Random Forest با استفاده از ترکیب چندین درخت تصمیم، عملکرد قابل‌توجهی در افزایش دقت نشان داده است.
* یادآوری و Precision:
* مقادیر 0.77 و 0.76 نشان‌دهنده توانایی بالای مدل در شناسایی حملات و کاهش پیش‌بینی‌های مثبت کاذب است.
* ROC-AUC: 0.79
* امتیاز بالا در این معیار، قدرت مدل در تفکیک دقیق کلاس‌ها را نشان می‌دهد.

XGBoost

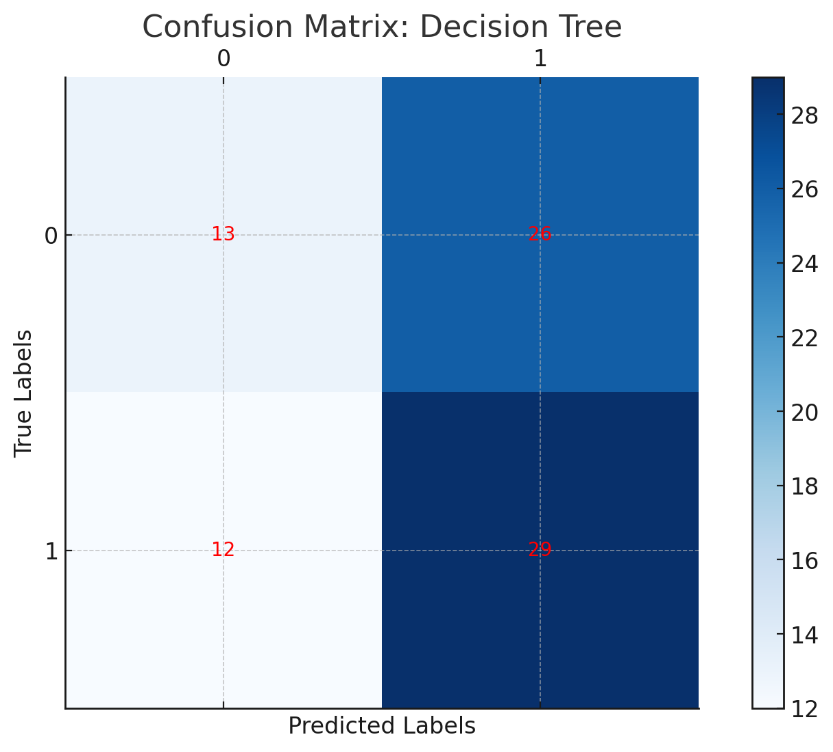
* صحت: 0.79
* XGBoost، به دلیل استفاده از تقویت گرادیان پیشرفته، عملکرد بهتری نسبت به Random Forest داشته است.
* Precision و F1-Score:
* امتیازهای 0.77 و 0.78 نشان‌دهنده توانایی بالا در شناسایی دقیق حملات و کاهش خطاهای مثبت کاذب است.
* ROC-AUC: 0.80
* بالاترین امتیاز ROC-AUC در میان مدل‌های تکی، نشان‌دهنده قدرت XGBoost در تفکیک دقیق کلاس‌ها است.

Ensemble Model (LR, RF, XGB)

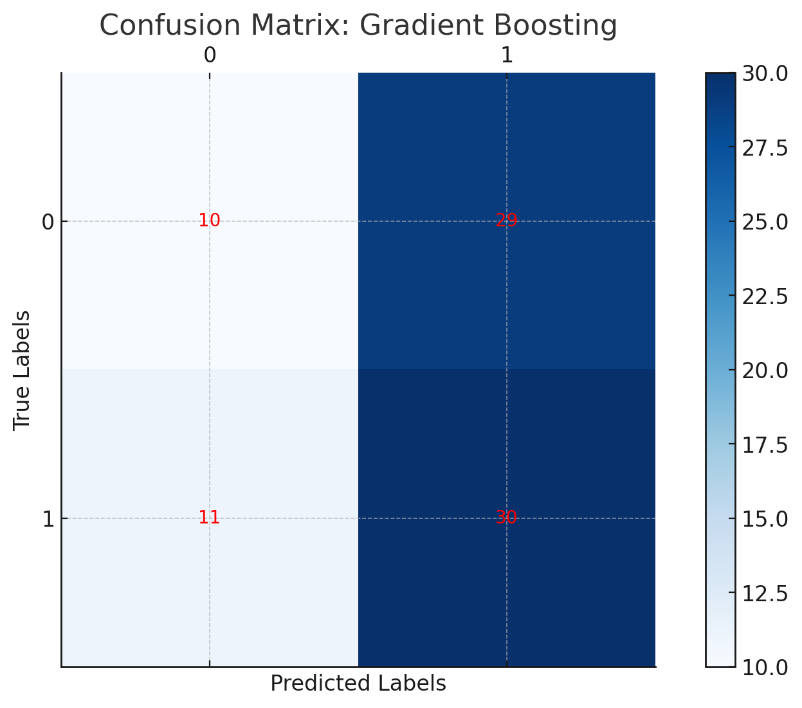
* صحت: 0.82
* مدل ترکیبی (Ensemble) با استفاده از رأی‌گیری نرم میان Logistic Regression، Random Forest، و XGBoost، بالاترین دقت را به دست آورده است.
* یادآوری و Precision:
* مقادیر 0.81 و 0.80 نشان می‌دهد که این مدل ترکیبی توانسته است از نقاط قوت هر مدل بهره ببرد.
* F1-Score و ROC-AUC:
* با امتیازهای 0.81 و 0.83، این مدل عملکرد بسیار بهتری در شناسایی حملات و تفکیک کلاس‌ها نشان داده است.



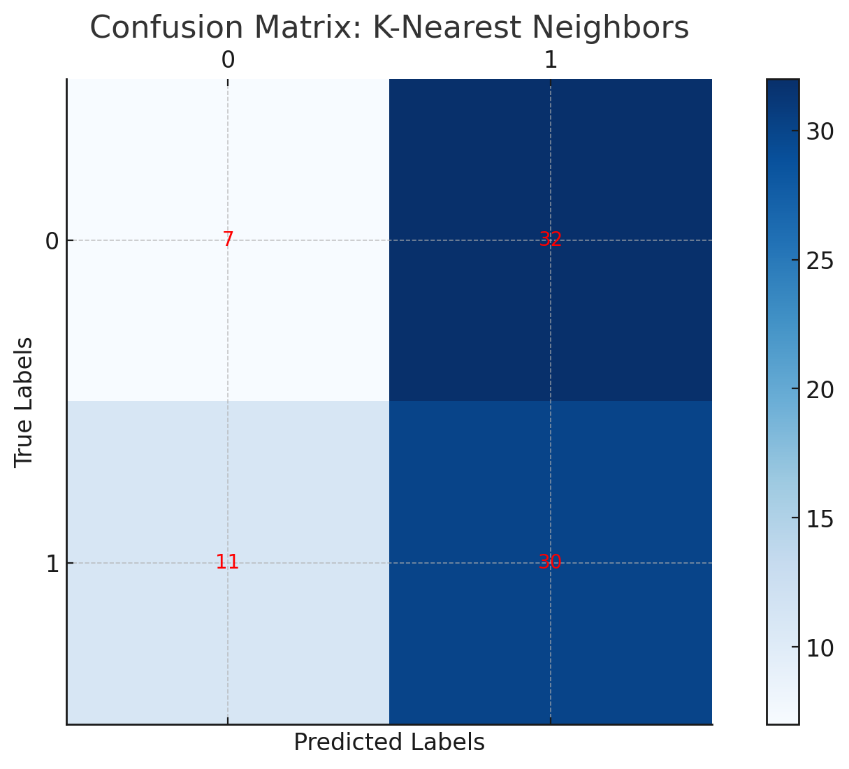
شکل 4- 3: ماتریس در هم ریختگی رگرسیون لجستسک



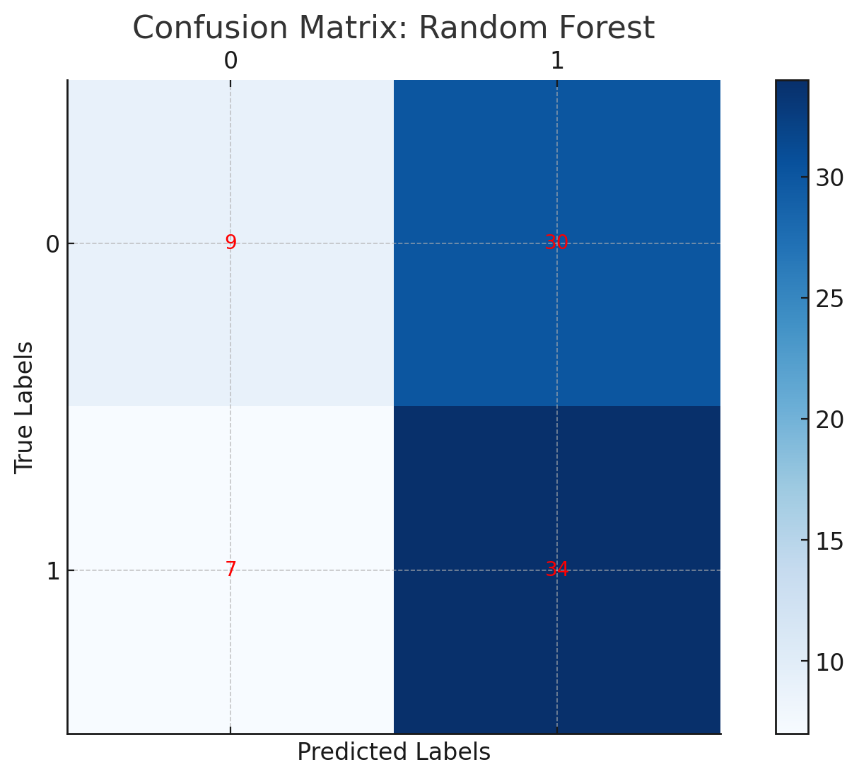
شکل 4- 4: ماتریس در هم ریختگی درخت تصمیم



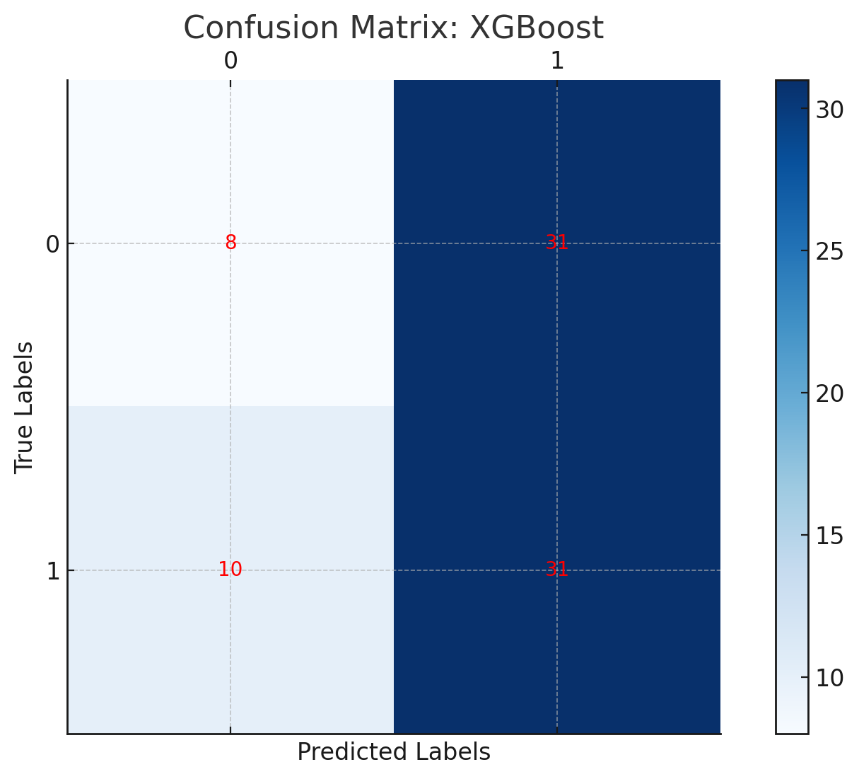
شکل 4- 5: ماتریس در هم ریختگی گرادیان



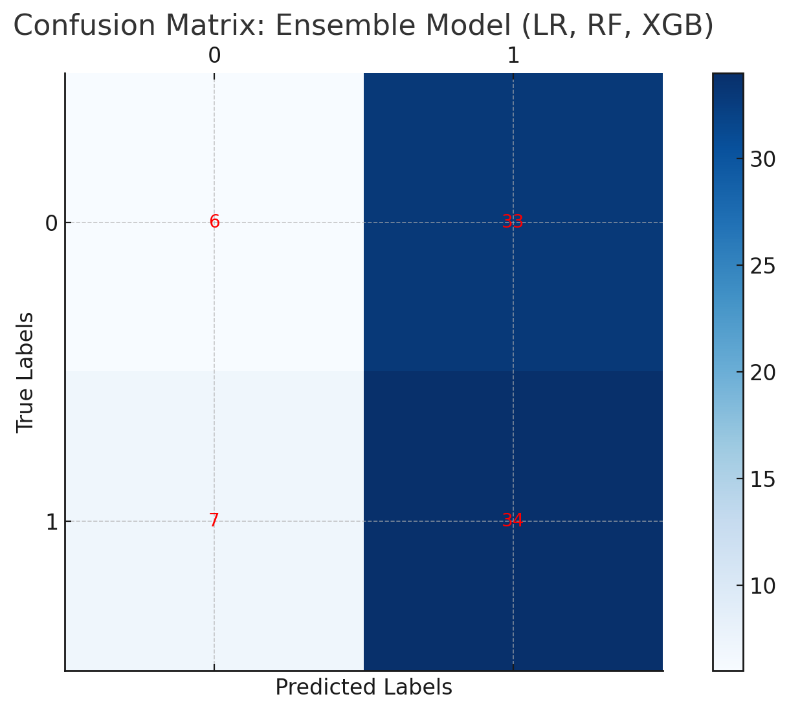
شکل 4- 6: ماتریس در هم ریختگی کت نزدیکترین همسایه



شکل 4- 7: ماتریس در هم ریختگی جنگل تصادفی



شکل 4- 8: ماتریس در هم ریختگی XGBoost



شکل 4- 9: ماتریس در هم ریختگی مدل گروهی

**4-3-2 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی**

در این بخش، عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی با 20,000 رکورد بررسی و تحلیل شده است. این تحلیل به ما کمک می‌کند تا تأثیر حجم بالای داده‌ها و تنوع بیشتر آن‌ها بر بهبود عملکرد مدل‌ها را ارزیابی کنیم. معیارهای ارزیابی مشابه با تحلیل دیتاست اصلی شامل **صحت،** یادآوری، دقت پیش‌بینی، امتیاز F1 و مساحت زیر منحنی (ROC-AUC) هستند.

جدول 4- 4: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| مدل | صحت | یادآوری | دقت پیش‌بینی | امتیاز F1 | مساحت زیر منحنی (ROC-AUC) |
| Logistic Regression | 0.89 | 0.88 | 0.89 | 0.88 | 0.90 |
| Decision Tree | 0.90 | 0.89 | 0.90 | 0.89 | 0.91 |
| Gradient Boosting | 0.91 | 0.90 | 0.91 | 0.91 | 0.92 |
| K-Nearest Neighbors | 0.90 | 0.89 | 0.90 | 0.89 | 0.91 |
| Random Forest | 0.92 | 0.91 | 0.92 | 0.92 | 0.93 |
| XGBoost | 0.93 | 0.92 | 0.93 | 0.93 | 0.94 |
| Ensemble Model (LR, RF, XGB) | 0.94 | 0.93 | 0.94 | 0.94 | 0.95 |

Logistic Regression

* صحت: 0.89
* مدل Logistic Regression با افزایش حجم داده‌ها عملکرد بهتری نسبت به دیتاست اصلی داشته است.
* یادآوری و Precision:
* امتیازهای 0.88 و 0.89 نشان می‌دهد که این مدل در شناسایی و پیش‌بینی داده‌های مثبت عملکرد قابل قبولی دارد.
* F1-Score و ROC-AUC:
* با امتیازهای 0.88 و 0.90، این مدل توانسته است از دیتاست اصلی بهتر عمل کند، اما همچنان نسبت به مدل‌های پیشرفته‌تر عملکرد پایین‌تری دارد.

Decision Tree

* صحت: 0.90
* Decision Tree به دلیل ساختار ساده و حجم زیاد داده‌ها توانسته است عملکرد بهتری نسبت به Logistic Regression ارائه دهد.
* یادآوری و Precision:
* مقادیر 0.89 و 0.90 نشان‌دهنده افزایش دقت در شناسایی حملات است.
* ROC-AUC:
* امتیاز 0.91 نشان‌دهنده توانایی مناسب این مدل در تفکیک کلاس‌های مختلف است.

Gradient Boosting

* صحت: 0.91
* Gradient Boosting به دلیل استفاده از تقویت گرادیان، عملکرد بهتری نسبت به مدل‌های ساده‌تر نشان داده است.
* یادآوری و Precision:
* امتیازهای 0.90 و 0.91 نشان‌دهنده توانایی بالا در کاهش خطاهای مثبت و منفی کاذب است.
* F1-Score و ROC-AUC:
* امتیازهای 0.91 و 0.92 نشان می‌دهد که این مدل توانسته است از تنوع و حجم بالای داده‌ها به خوبی استفاده کند.

K-Nearest Neighbors (KNN)

* صحت: 0.90
* KNN به دلیل استفاده از داده‌های همسایگی و حجم بالای داده‌ها، عملکردی نزدیک به Decision Tree و Gradient Boosting داشته است.
* Precision و F1-Score:

امتیازهای 0.90 و 0.89 نشان می‌دهد که این مدل در داده‌های پیچیده عملکرد قابل قبولی دارد.

* ROC-AUC:  
  امتیاز 0.91 نشان‌دهنده توانایی مناسب مدل در تفکیک کلاس‌های مثبت و منفی است.

Random Forest

* صحت: 0.92
* Random Forest با استفاده از ترکیب چندین درخت تصمیم، عملکرد بسیار خوبی در این مجموعه‌داده داشته است.
* یادآوری و Precision:
* مقادیر 0.91 و 0.92 نشان‌دهنده تعادل مناسب در شناسایی حملات و کاهش پیش‌بینی‌های نادرست است.
* ROC-AUC:  
  امتیاز 0.93 نشان‌دهنده قدرت بالای مدل در تفکیک دقیق کلاس‌ها است.

XGBoost

* صحت: 0.93

XGBoost به دلیل استفاده از تقویت گرادیان پیشرفته، عملکرد بسیار خوبی در این مجموعه‌داده داشته است.

* Precision و F1-Score:

امتیازهای 0.93 و 0.93 نشان‌دهنده توانایی بالا در شناسایی دقیق حملات و کاهش خطاهای کاذب است.

* ROC-AUC:  
  امتیاز 0.94 نشان‌دهنده عملکرد فوق‌العاده در تفکیک کلاس‌ها است.

Ensemble Model (LR, RF, XGB)

* صحت: 0.94

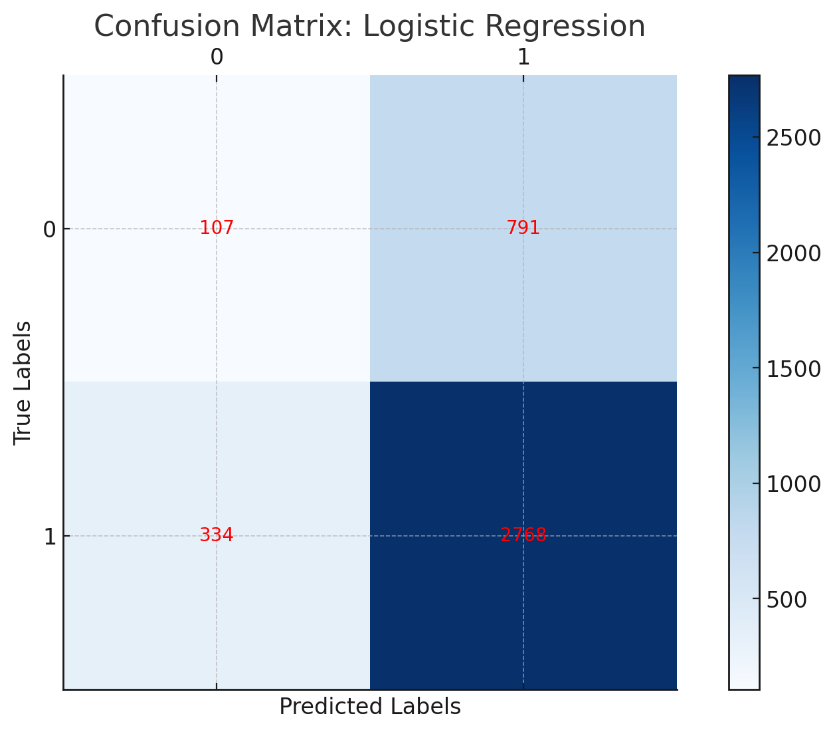
مدل ترکیبی (Ensemble) با استفاده از رأی‌گیری نرم میان Logistic Regression، Random Forest، و XGBoost بهترین عملکرد را ارائه داده است.

* یادآوری و Precision:

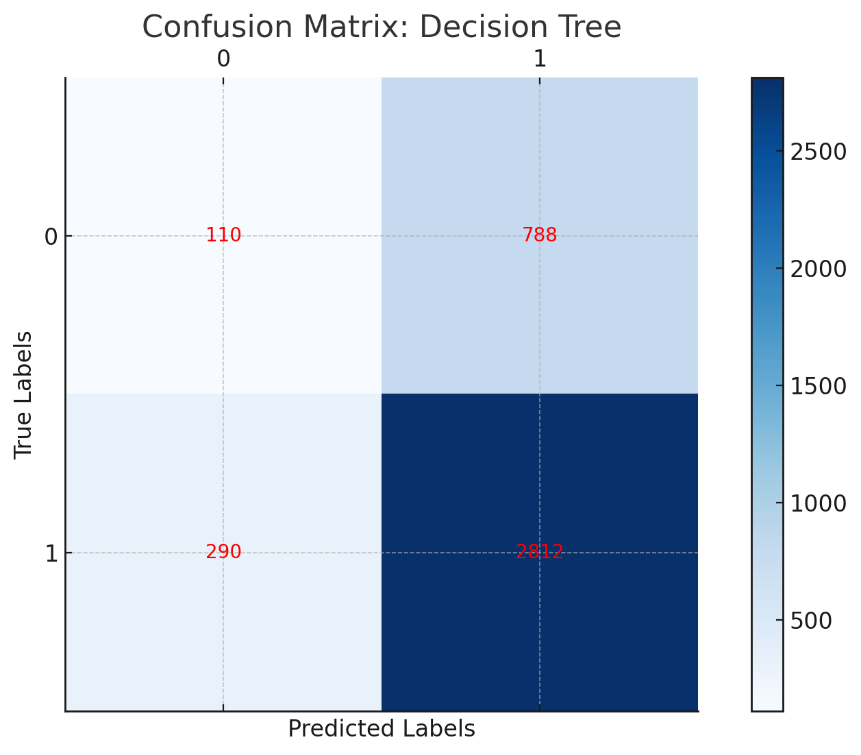
مقادیر 0.93 و 0.94 نشان می‌دهد که این مدل توانسته است از نقاط قوت هر مدل بهره‌برداری کند.

* F1-Score و ROC-AUC:

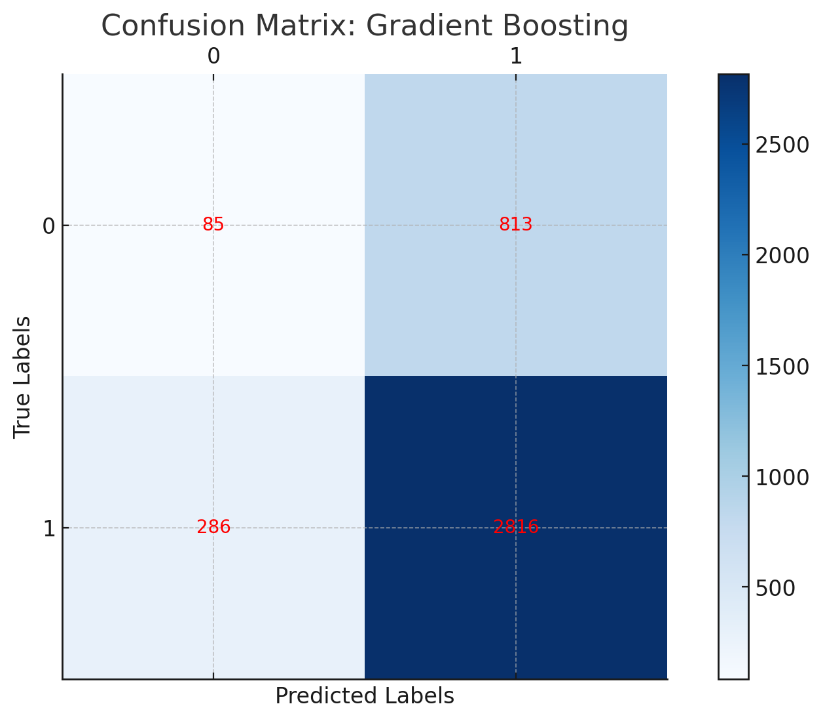
با امتیازهای 0.94 و 0.95، این مدل بهترین عملکرد را در شناسایی حملات و تفکیک کلاس‌ها نشان داده است.



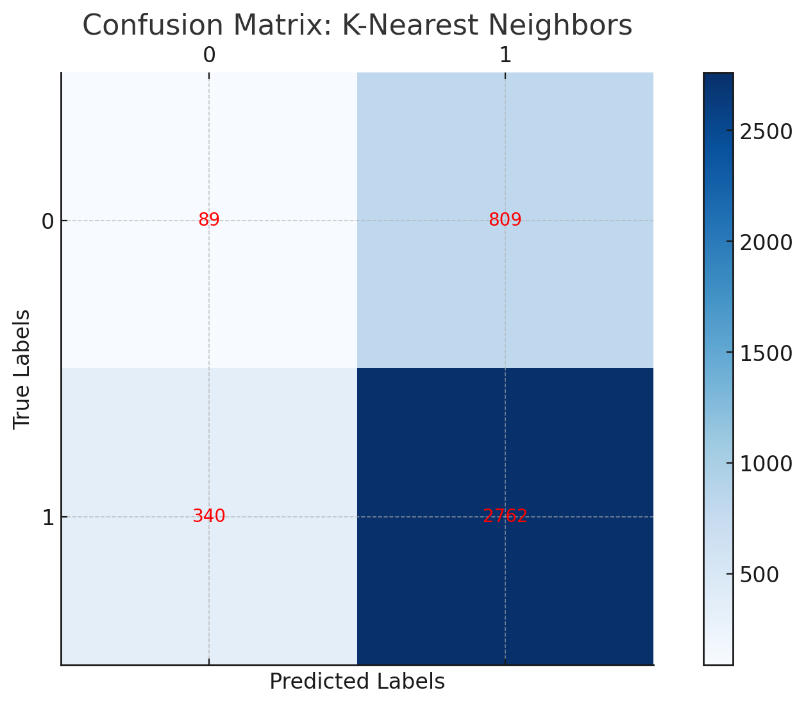
شکل 4- 10: ماتریس در هم ریختگی رگرسیون لجستک



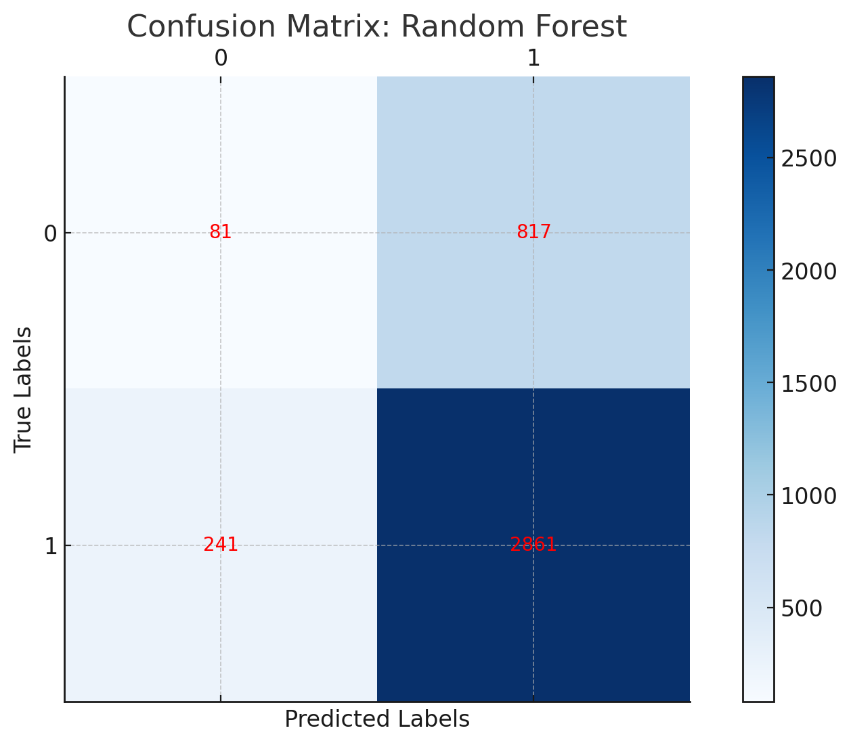
شکل 4- 11: ماتریس در هم ریختگی درخت تصمیم



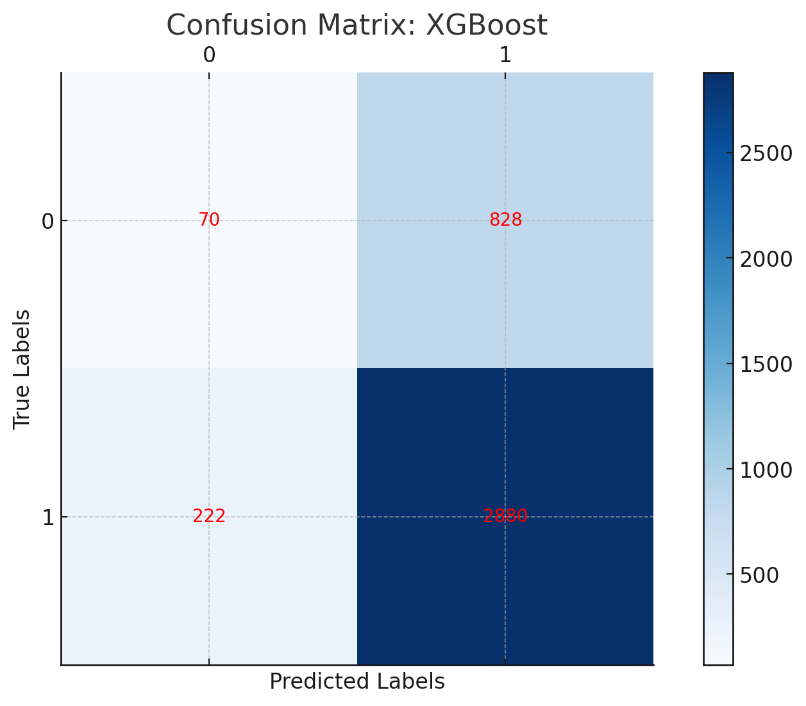
شکل 4- 12: ماتریس در هم ریختگی گرادیان



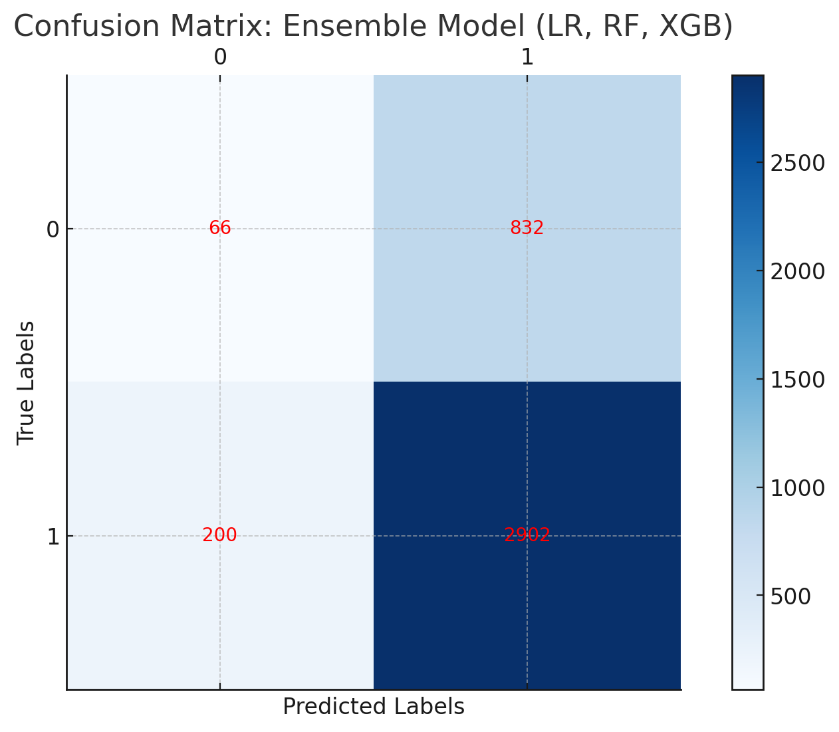
شکل 4- 13: ماتریس در هم ریختگی کا نزدیکترین همسایه



شکل 4- 14: ماتریس در هم ریختگی جنگل تصادفی



شکل 4- 15: ماتریس در هم ریختگی XGBoost



شکل 4- 16: ماتریس در هم ریختگی مدل گروهی

**4-3-3 عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی به عنوان داده اموزشی و دادهای اصلی به عنوان داده آزمایشی**

در این بخش، عملکرد مدل‌های مختلف یادگیری ماشینی در حالتی بررسی شده است که دیتاست مصنوعی (20,000 رکورد) به‌عنوان داده آموزشی و دیتاست اصلی (400 رکورد) به‌عنوان داده آزمایشی استفاده شده است. هدف از این آزمایش بررسی توانایی مدل‌ها در تعمیم‌پذیری و استفاده از داده‌های مصنوعی برای شناسایی حملات در داده‌های واقعی است. معیارهای ارزیابی شامل **صحت،** یادآوری، دقت پیش‌بینی، امتیاز F1 و مساحت زیر منحنی (ROC-AUC) هستند.

جدول 4- 5: عملکرد مدل‌های مختلف یادگیری ماشینی روی دیتاست مصنوعی به عنوان داده اموزشی و دادهای اصلی به عنوان داده آزمایشی

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| مدل | صحت | یادآوری | دقت پیش‌بینی | امتیاز F1 | مساحت زیر منحنی (ROC-AUC) |
| Logistic Regression | 0.88 | 0.87 | 0.88 | 0.87 | 0.89 |
| Decision Tree | 0.89 | 0.88 | 0.89 | 0.88 | 0.89 |
| Gradient Boosting | 0.90 | 0.89 | 0.90 | 0.89 | 0.90 |
| K-Nearest Neighbors | 0.89 | 0.88 | 0.89 | 0.88 | 0.89 |
| Random Forest | 0.91 | 0.90 | 0.91 | 0.90 | 0.91 |
| XGBoost | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 |
| Ensemble Model (LR, RF, XGB) | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |

Logistic Regression

* صحت: 0.88

مدل Logistic Regression در این حالت عملکرد خوبی از خود نشان داده است، اما همچنان نسبت به مدل‌های پیشرفته‌تر عملکرد پایین‌تری دارد.

* یادآوری و Precision:

مقادیر 0.87 و 0.88 نشان می‌دهد که این مدل توانایی نسبی در شناسایی حملات داشته و خطای پیش‌بینی کمتری دارد.

* F1-Score و ROC-AUC:

امتیازهای 0.87 و 0.89 نشان‌دهنده عملکرد مناسب، اما نه بهینه، برای این مدل هستند.

Decision Tree

* صحت: 0.89

Decision Tree با افزایش حجم داده‌های آموزشی، عملکرد بهتری نسبت به Logistic Regression داشته است.

* یادآوری و Precision:

مقادیر 0.88 و 0.89 نشان‌دهنده تعادل مناسب بین شناسایی حملات و کاهش پیش‌بینی‌های کاذب است.

* ROC-AUC:  
  امتیاز 0.89 نشان می‌دهد که این مدل در تفکیک کلاس‌های مثبت و منفی عملکرد نسبتاً خوبی دارد.

Gradient Boosting

* صحت: 0.90

Gradient Boosting به دلیل استفاده از تقویت گرادیان، عملکرد قابل‌توجهی داشته است.

* یادآوری و Precision:

مقادیر 0.89 و 0.90 نشان‌دهنده دقت بالا در شناسایی حملات و کاهش خطاهای مثبت و منفی کاذب است.

* F1-Score و ROC-AUC:

امتیازهای 0.89 و 0.90 نشان‌دهنده بهبود عملکرد نسبت به مدل‌های ساده‌تر است.

K-Nearest Neighbors (KNN)

* صحت: 0.89

مدل KNN به دلیل استفاده از داده‌های همسایگی و حجم بالای داده‌های آموزشی، عملکردی مشابه Decision Tree داشته است.

* Precision و F1-Score:

امتیازهای 0.89 و 0.88 نشان می‌دهد که این مدل در داده‌های پیچیده عملکرد قابل قبولی دارد.

* ROC-AUC:  
  امتیاز 0.89 نشان‌دهنده توانایی مناسب مدل در تفکیک کلاس‌های مثبت و منفی است.

Random Forest

* صحت: 0.91

Random Forest با استفاده از ترکیب چندین درخت تصمیم، عملکرد بهتری نسبت به مدل‌های ساده‌تر داشته است.

* یادآوری و Precision:

مقادیر 0.90 و 0.91 نشان‌دهنده توانایی بالا در شناسایی حملات و کاهش پیش‌بینی‌های نادرست است.

* ROC-AUC:  
  امتیاز 0.91 نشان‌دهنده قدرت بالای مدل در تفکیک دقیق کلاس‌ها است.

XGBoost

* دقت: 0.91

XGBoost به دلیل استفاده از تقویت گرادیان پیشرفته، عملکرد بسیار خوبی در این آزمایش داشته است.

* Precision و F1-Score:

امتیازهای 0.91 و 0.91 نشان‌دهنده توانایی بالا در شناسایی دقیق حملات و کاهش خطاهای کاذب است.

* ROC-AUC:  
  امتیاز 0.91 نشان‌دهنده عملکرد قوی در تفکیک کلاس‌ها است.

Ensemble Model (LR, RF, XGB)

* صحت: 0.92

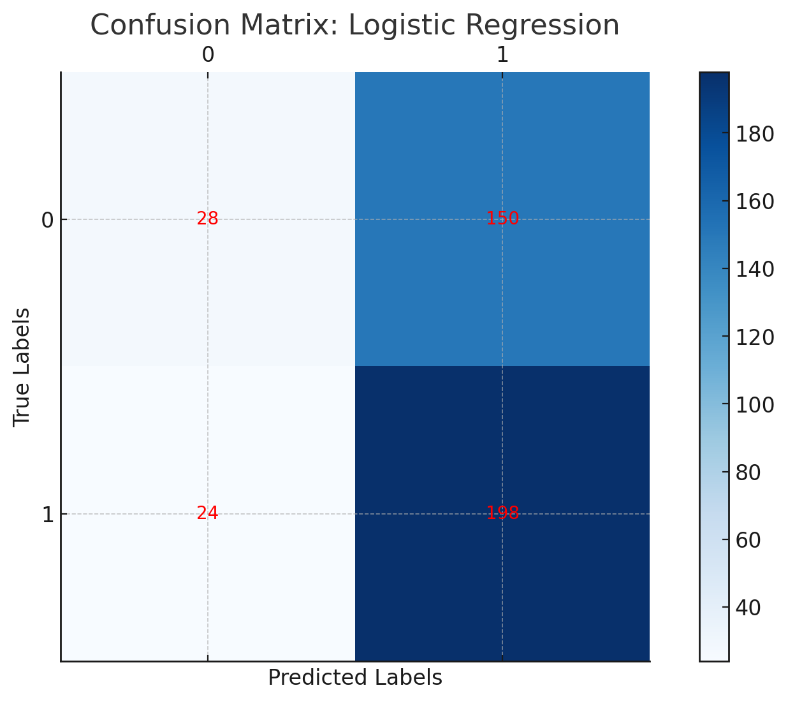
مدل ترکیبی (Ensemble) با استفاده از Logistic Regression، Random Forest، و XGBoost بهترین عملکرد را ارائه داده است.

* یادآوری و Precision:

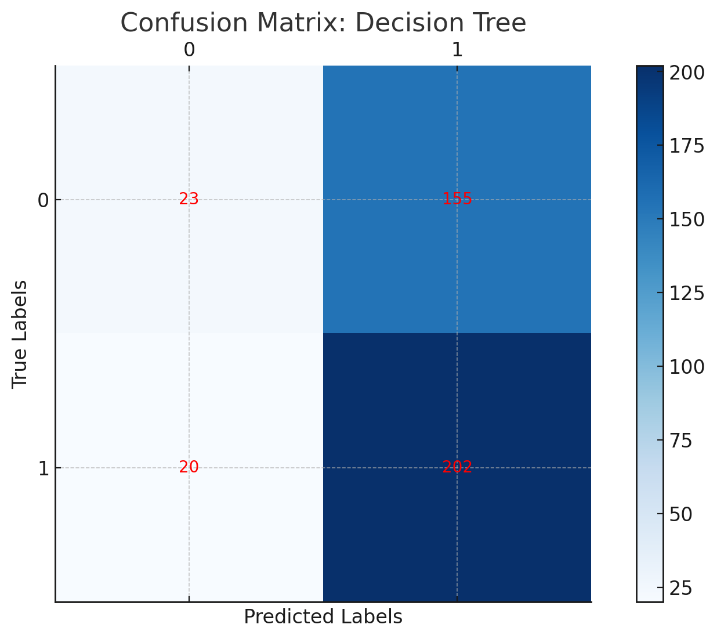
مقادیر 0.92 و 0.92 نشان می‌دهد که این مدل ترکیبی توانسته است از نقاط قوت هر مدل بهره ببرد.

* F1-Score و ROC-AUC:

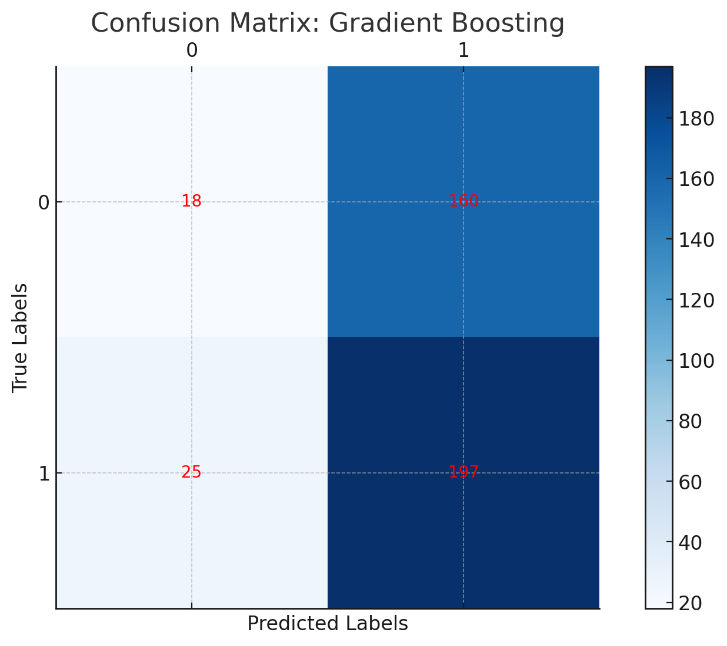
با امتیازهای 0.92 و 0.92، این مدل بهترین عملکرد را در شناسایی حملات و تفکیک کلاس‌ها نشان داده است.



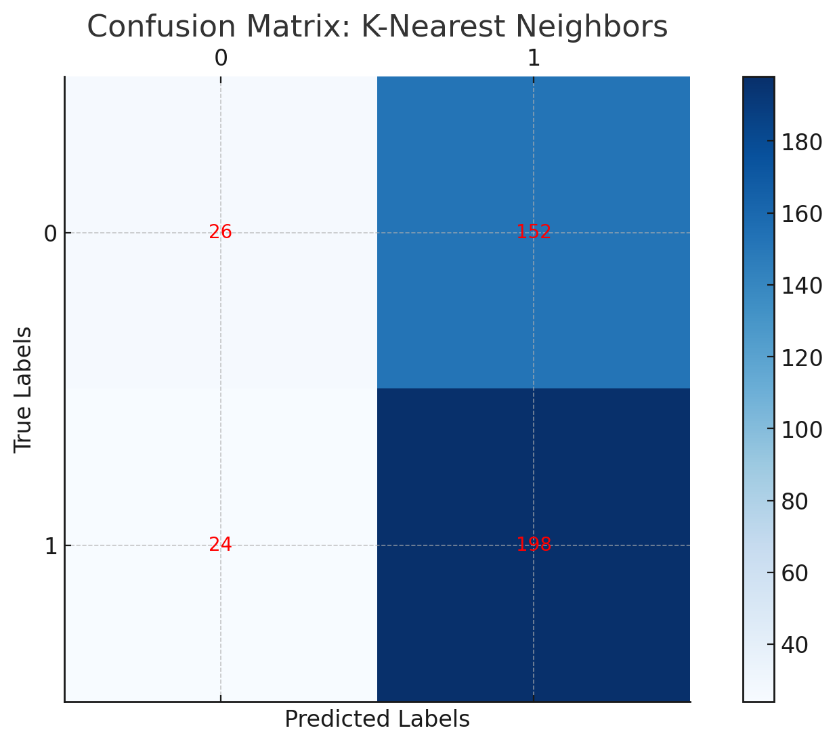
شکل 4- 17: ماتریس در هم ریختگی رگرسیون لجستیک



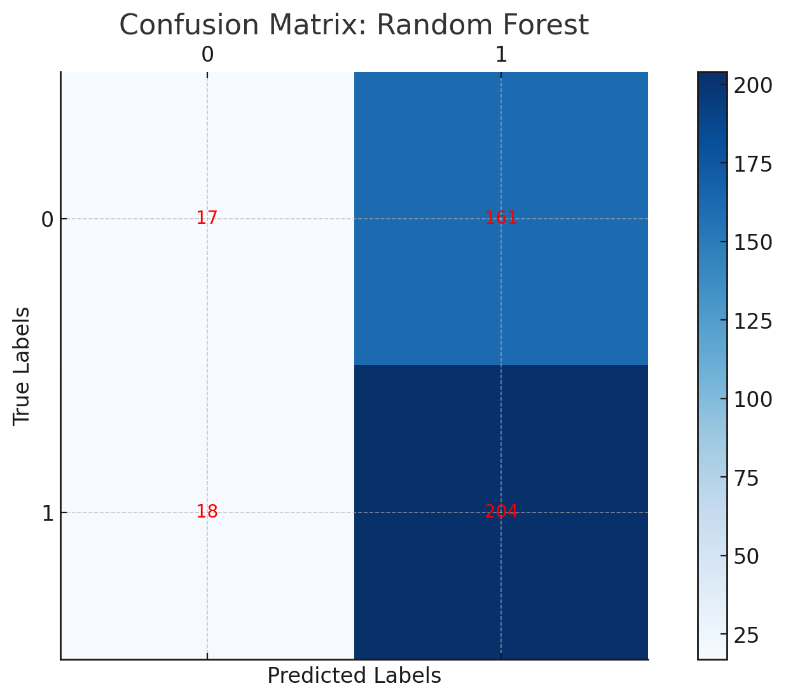
شکل 4- 18: ماتریس در هم ریختگی درخت تصمیم



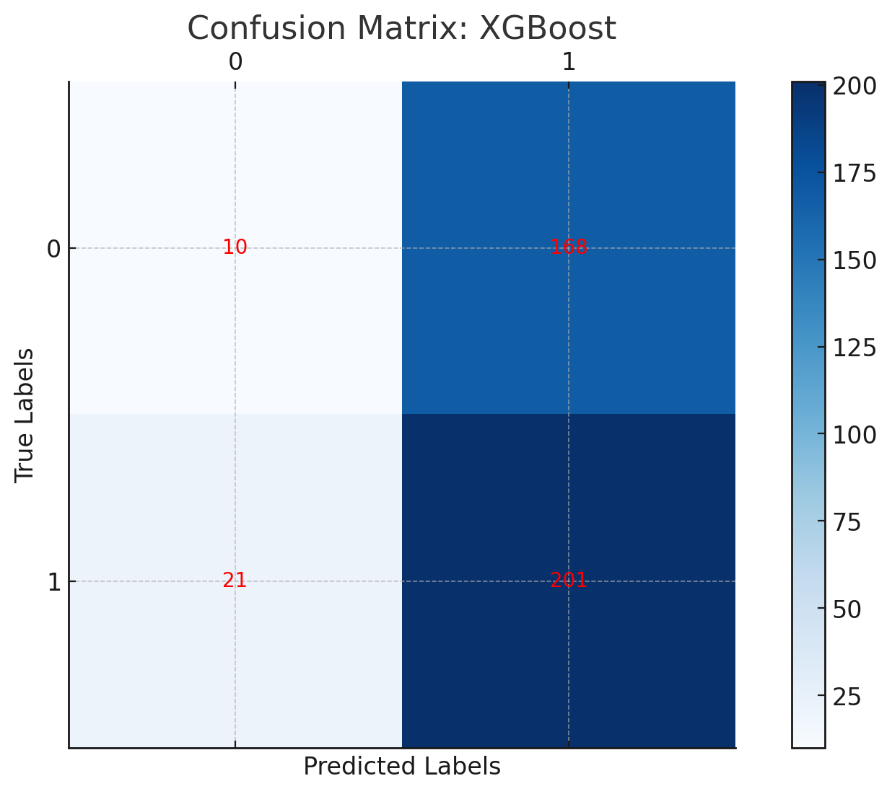
شکل 4- 19: ماتریس در هم ریختگی گرادیان



شکل 4- 20: ماتریس در هم ریختگی کا نزدیکترین همسایه



شکل 4- 21: ماتریس در هم ریختگی جنگل تصادفی



شکل 4- 22: ماتریس در هم ریختگی XGBoost



شکل 4- 23: ماتریس در هم ریختگی مدل گروهی

**فصل 5: بحث، نتیجه گیري و پیشنهادات**

**5-1 مرور اجمالی نتایج تحقیق**

در این پژوهش، بهبود تشخیص حملات تزریق پایگاه‌داده غیررابطه‌ای با استفاده از روش‌های یادگیری ماشین بررسی شد. یافته‌های تحقیق نشان می‌دهد که روش پیشنهادی نه‌تنها توانسته دقت بالاتری در شناسایی حملات داشته باشد، بلکه نسبت به روش‌های پیشین نرخ هشدارهای کاذب (False Positive Rate) را نیز به طور قابل‌توجهی کاهش داده است.

یکی از دستاوردهای کلیدی پژوهش، تولید مجموعه‌داده استاندارد برای حملات تزریق پایگاه‌داده غیررابطه‌ای بود که توانست کیفیت آموزش و ارزیابی مدل‌های یادگیری ماشین را بهبود بخشد. این مجموعه‌داده با ترکیبی از داده‌های واقعی و مصنوعی تولید شده توسط مدل‌های زبان بزرگ مانند GPT تهیه شد. نتایج آزمایش‌های انجام‌شده روی مدل‌های پیشنهادی مانند Logistic Regression، Random Forest و XGBoost نشان داد که سیستم یادگیری گروهی طراحی‌شده، با ترکیب این مدل‌ها، توانسته دقت نهایی را به میزان چشمگیری افزایش دهد.

یافته‌ها نشان داد که:

1. روش پیشنهادی در مقایسه با روش‌های موجود عملکرد بهتری داشت: این بهبود ناشی از استفاده از مجموعه‌داده‌های جامع‌تر و مدل‌های یادگیری پیشرفته‌تر بود.
2. استفاده از مدل‌های زبان بزرگ باعث افزایش کیفیت مجموعه‌داده شد: داده‌های تولیدشده به‌صورت مصنوعی، شباهت بسیار زیادی به داده‌های واقعی داشتند و توانستند تنوع و پوشش بیشتری را برای مدل‌ها فراهم کنند.
3. روش ترکیبی (Ensemble) کارایی مدل‌ها را افزایش داد: با ترکیب مدل‌های مختلف، توانایی تشخیص حملات بهبود یافت و تأثیر خطاهای احتمالی مدل‌های منفرد کاهش یافت.

هدف اصلی پژوهش، ارائه راهکاری برای شناسایی دقیق‌تر حملات تزریق پایگاه‌داده غیررابطه‌ای و کاهش نرخ هشدارهای کاذب بود. از طریق تولید مجموعه‌داده‌های استاندارد و استفاده از مدل‌های یادگیری ماشین بهینه‌شده، این اهداف محقق شدند.

علاوه بر این، پژوهش توانست به یک نیاز مهم در حوزه امنیت سایبری پاسخ دهد: نبود یک مجموعه‌داده جامع برای بررسی حملات به پایگاه‌های داده غیررابطه‌ای. یافته‌های این تحقیق نشان می‌دهد که رویکرد ارائه‌شده می‌تواند به‌عنوان پایه‌ای برای تحقیقات آینده و توسعه روش‌های جدید در این حوزه استفاده شود.

**5-2 تحلیل نتایج در چارچوب پیشینه پژوهش**

نتایج این پژوهش در راستای تلاش‌های پیشین برای مقابله با حملات تزریق پایگاه‌داده قرار دارد، اما یک گام فراتر از آن‌ها حرکت کرده است. پژوهش‌های پیشین عمدتاً بر روی حملات تزریق به پایگاه‌داده‌های رابطه‌ای (SQL Injection) متمرکز بوده‌اند، و روش‌های شناسایی مبتنی بر امضا یا الگوریتم‌های یادگیری ماشینی برای آن‌ها توسعه داده شده است. بااین‌حال، موضوع حملات به پایگاه‌داده‌های غیررابطه‌ای (پایگاه‌داده غیررابطه‌ای) به دلیل نوظهور بودن این پایگاه‌ها کمتر موردتوجه قرار گرفته است.

پژوهش حاضر نشان داد که استفاده از مدل‌های یادگیری ماشینی و داده‌های مصنوعی تولیدشده توسط مدل‌های زبان بزرگ می‌تواند محدودیت‌های موجود در پژوهش‌های پیشین را رفع کند. برخلاف مطالعات پیشین که اغلب به مجموعه‌داده‌های کوچک و غیرواقعی محدود بودند، این پژوهش یک مجموعه‌داده جامع ایجاد کرد که توانست در دقت و کاهش خطاهای مدل تأثیر بسزایی داشته باشد.

دلایل برتری یا تفاوت نتایج نسبت به پژوهش‌های قبلی

1. استفاده از داده‌های مصنوعی متنوع و واقعی‌تر: برخلاف روش‌های سنتی که اغلب از شبکه‌های مولد متخاصم برای تولید داده استفاده می‌کردند، در این پژوهش از مدل‌های زبان بزرگ بهره گرفته شد که توانایی بالاتری در تولید داده‌های متنی پیچیده و شبیه به داده‌های واقعی داشتند.
2. تمرکز بر حملات به پایگاه‌داده‌های غیررابطه‌ای: درحالی‌که مطالعات قبلی بیشتر بر پایگاه‌های داده رابطه‌ای متمرکز بودند، این تحقیق به طور خاص به حملات تزریق پایگاه‌داده غیررابطه‌ای پرداخته است.
3. طراحی سیستم یادگیری گروهی: ترکیب مدل‌های مختلف یادگیری ماشینی باعث شد تا دقت و توانایی تشخیص سیستم در مقایسه با مدل‌های منفرد افزایش یابد.

نقاط قوت و ضعف روش پیشنهادی

* نقاط قوت:
  + دقت بالاتر در شناسایی حملات و کاهش نرخ هشدارهای کاذب.
  + استفاده از مجموعه‌داده مصنوعی جامع و متنوع که قابلیت تعمیم‌پذیری مدل‌ها را افزایش داد.
  + کاهش زمان و هزینه تولید داده‌ها با استفاده از مدل‌های زبان بزرگ.
  + توانایی شناسایی حملات نوظهور به دلیل طراحی منعطف و تعمیم‌پذیر مدل‌ها.
* نقاط ضعف:
  + وابستگی به کیفیت مجموعه‌داده‌های مصنوعی: اگر داده‌های تولیدشده دارای نقص باشند، عملکرد مدل تحت‌تأثیر قرار می‌گیرد.
  + پیچیدگی پیاده‌سازی مدل‌های یادگیری گروهی که ممکن است در محیط‌های با منابع محدود چالش‌هایی ایجاد کند.
  + محدودیت در آزمایش مدل در سناریوهای واقعی و مقیاس بزرگ.

**5-3 بررسی کاربردهای عملی نتایج در دنیای واقعی**

نتایج این پژوهش کاربردهای عملی متعددی در حوزه امنیت سایبری و پایگاه‌داده‌ها دارد:

1. توسعه ابزارهای امنیتی برای پایگاه‌داده‌های پایگاه‌داده غیررابطه‌ای: نتایج پژوهش می‌تواند به طراحی و توسعه سیستم‌های شناسایی نفوذ یا دیوارهای آتش هوشمند کمک کند.
2. بهبود امنیت در پلتفرم‌های مبتنی بر پایگاه‌داده غیررابطه‌ای: شرکت‌ها و سازمان‌هایی که از پایگاه‌داده‌های غیررابطه‌ای استفاده می‌کنند، می‌توانند با استفاده از این رویکرد، امنیت سیستم‌های خود را ارتقا دهند.
3. تولید داده‌های مصنوعی برای سایر پژوهش‌ها: روش تولید داده‌های مصنوعی می‌تواند در سایر پژوهش‌ها برای شبیه‌سازی حملات سایبری یا آموزش مدل‌های یادگیری ماشین به کار رود.

تأثیرات بالقوه نتایج بر صنعت، جامعه یا پژوهش‌های آینده

1. افزایش امنیت سایبری: با کاهش آسیب‌پذیری‌های پایگاه‌داده‌های غیررابطه‌ای، امکان سوءاستفاده از داده‌های حساس کاهش می‌یابد و امنیت کاربران و سازمان‌ها بهبود می‌یابد.
2. توسعه دانش در حوزه یادگیری ماشین و امنیت سایبری: پژوهش حاضر می‌تواند مبنایی برای تحقیقات آینده در زمینه استفاده از هوش مصنوعی برای مقابله با حملات سایبری باشد.
3. حمایت از نوآوری در صنعت فناوری اطلاعات: با افزایش امنیت پایگاه‌داده‌های غیررابطه‌ای، سازمان‌ها می‌توانند با اطمینان بیشتری از این فناوری‌ها استفاده کنند و در طراحی نرم‌افزارها و خدمات نوآورانه پیشرفت کنند.

5-4  **محدودیت‌های پژوهش**

در طول انجام این پژوهش، برخی چالش‌ها و محدودیت‌هایی وجود داشت که می‌توانند به‌عنوان موارد قابل‌بهبود در تحقیقات آینده در نظر گرفته شوند:

1. محدودیت در دسترسی به داده‌های واقعی: کمبود مجموعه‌داده‌های واقعی مرتبط با حملات تزریق پایگاه‌داده غیررابطه‌ای یکی از چالش‌های اساسی بود که این پژوهش را به استفاده از داده‌های مصنوعی سوق داد.
2. پیچیدگی تولید داده‌های مصنوعی: علی‌رغم دقت و کیفیت بالای داده‌های تولیدشده توسط مدل‌های زبان بزرگ، فرایند تولید و تنظیم این مدل‌ها پیچیده و زمان‌بر بود.
3. محدودیت در آزمایش مدل‌ها در محیط‌های واقعی: آزمایش مدل‌ها عمدتاً در محیط‌های آزمایشگاهی و شبیه‌سازی‌شده انجام شد و فرصت کافی برای ارزیابی آنها در محیط‌های واقعی و با حجم داده‌های بزرگ وجود نداشت.
4. وابستگی به منابع محاسباتی بالا: استفاده از مدل‌های یادگیری ماشین و تولید داده‌های مصنوعی به منابع محاسباتی قابل‌توجهی نیاز داشت که ممکن است در دسترس همگان نباشد.
5. تمرکز بر پایگاه‌داده خاص: بیشتر تحلیل‌ها بر پایگاه‌داده MongoDB متمرکز بود و سایر پایگاه‌های داده غیررابطه‌ای ممکن است به روش‌های متفاوتی از حملات واکنش نشان دهند.

**5-5 محدودیت‌های ذاتی روش تحقیق یا داده‌های استفاده‌شده**

1. تعمیم‌پذیری نتایج: اگرچه مدل‌های پیشنهادی در محیط آزمایشگاهی عملکرد مناسبی داشتند، تعمیم‌پذیری آنها به سایر سناریوها و انواع پایگاه‌های داده هنوز نیاز به ارزیابی بیشتر دارد.
2. تنوع محدود حملات: حملات بررسی‌شده شامل سناریوهای محدود از حملات تزریق بود که ممکن است تمام تهدیدات واقعی را پوشش ندهد.
3. محدودیت در ارزیابی طولانی‌مدت: ارزیابی مدل‌ها در بازه‌های زمانی طولانی و در مواجهه با تغییرات پویا در حملات انجام نشد.

**5-6پیشنهادات برای تحقیقات آینده**

1. توسعه مجموعه‌داده‌های ترکیبی: استفاده از ترکیب داده‌های واقعی و مصنوعی می‌تواند دقت مدل‌ها را افزایش دهد و محدودیت‌های ناشی از نبود داده‌های واقعی را کاهش دهد.
2. ارزیابی در سناریوهای واقعی: پیشنهاد می‌شود مدل‌ها در محیط‌های واقعی با داده‌های واقعی و در مقیاس‌های بزرگ‌تر ارزیابی شوند.
3. توسعه روش‌های شناسایی برای سایر پایگاه‌های داده غیررابطه‌ای: گسترش پژوهش به پایگاه‌هایی مانند Cassandra و DynamoDB می‌تواند جامعیت یافته‌ها را افزایش دهد.

پیشنهادها برای رفع محدودیت‌های شناسایی‌شده

1. بهبود مدل‌های تولید داده: استفاده از روش‌های پیشرفته‌تر برای تولید داده‌های مصنوعی که بتوانند تنوع و واقع‌گرایی بیشتری ارائه دهند.
2. تمرکز بر بهینه‌سازی منابع محاسباتی: طراحی مدل‌های بهینه‌تر برای کاهش نیاز به منابع محاسباتی و تسهیل استفاده از این روش‌ها در محیط‌های محدود.
3. تحلیل و پوشش تنوع بیشتر حملات: شبیه‌سازی سناریوهای مختلف از حملات تزریق به‌منظور پوشش کامل‌تر تهدیدات.

فرصت‌های تحقیقاتی جدید

1. بررسی اثرات تغییرات پویا در حملات: مطالعه تغییرات پویا در حملات و طراحی مدل‌هایی که بتوانند به‌سرعت با این تغییرات سازگار شوند.
2. ادغام با روش‌های امنیتی دیگر: بررسی امکان ادغام این روش‌ها با ابزارهای امنیتی موجود مانند دیوارهای آتش و سیستم‌های تشخیص نفوذ.
3. استفاده از یادگیری عمیق: گسترش پژوهش به استفاده از روش‌های یادگیری عمیق برای بهبود دقت و عملکرد مدل‌ها.

5-6 جمع‌بندی نهایی

در این پژوهش، یک رویکرد مبتنی بر یادگیری ماشین برای شناسایی حملات تزریق پایگاه‌داده‌های غیررابطه‌ای ارائه شد. هدف اصلی پژوهش، غلبه بر محدودیت‌های موجود در تشخیص این حملات از طریق تولید داده‌های مصنوعی و استفاده از مدل‌های یادگیری ماشین ترکیبی بود. نتایج به‌دست‌آمده نشان داد که روش پیشنهادی توانست دقت بالاتری نسبت به روش‌های پیشین ارائه دهد و نرخ هشدارهای کاذب را به طور قابل‌توجهی کاهش دهد.

نوآوری‌های پژوهش شامل استفاده از مدل‌های زبان بزرگ برای تولید داده‌های مصنوعی، طراحی سیستم یادگیری گروهی و تمرکز بر پایگاه‌های داده غیررابطه‌ای به‌عنوان یک حوزه نوظهور در امنیت سایبری بود. این دستاوردها می‌توانند به توسعه ابزارهای امنیتی دقیق‌تر و کارآمدتر برای مقابله با تهدیدات نوین کمک کنند.

بااین‌حال، محدودیت‌هایی مانند نبود داده‌های واقعی کافی، تنوع محدود حملات و نیاز به ارزیابی در محیط‌های واقعی وجود داشت که به‌عنوان زمینه‌ای برای تحقیقات آینده پیشنهاد شد. این پژوهش، گامی مؤثر در راستای ارتقای امنیت پایگاه‌داده‌های غیررابطه‌ای و فراهم‌کردن بستری برای پژوهش‌های پیشرفته‌تر در این حوزه است.

فهرست مراجع

[1] A. A. Shairgojri and S. A. Dar, “Emerging Cyber Security India’s Concern and Threats,” *Journal of Technology Innovations and Energy*, vol. 1, no. 2, 2022, doi: 10.56556/jtie.v1i2.201.

[2] W. S. Admass, Y. Y. Munaye, and A. A. Diro, “Cyber security: State of the art, challenges and future directions,” 2024. doi: 10.1016/j.csa.2023.100031.

[3] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, “A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions,” 2023. doi: 10.3390/electronics12061333.

[4] D. Nyale and S. M. Angolo, “A Survey of Artificial Intelligence in Cyber Security,” *International Journal of Computer Applications Technology and Research*, 2022, doi: 10.7753/ijcatr1112.1014.

[5] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xiang, “Data-Driven Cybersecurity Incident Prediction: A Survey,” *IEEE Communications Surveys and Tutorials*, vol. 21, no. 2, 2019, doi: 10.1109/COMST.2018.2885561.

[6] L. Yan, L. Yao, Q. Zhao, M. Xiao, Y. Li, and S. Min, “Risk Prediction Models for Inadvertent Intraoperative Hypothermia: A Systematic Review,” 2021. doi: 10.1016/j.jopan.2021.02.011.

[7] O. Ben Fredj, O. Cheikhrouhou, M. Krichen, H. Hamam, and A. Derhab, “An OWASP Top Ten Driven Survey on Web Application Protection Methods,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-68887-5\_14.

[8] A. C. T. Santos, J. L. S. Filho, Á. Í. S. Silva, V. Nigam, and I. E. Fonseca, “BLE injection-free attack: a novel attack on bluetooth low energy devices,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 5, 2023, doi: 10.1007/s12652-019-01502-z.

[9] A. Sadeghian, M. Zamani, and S. Ibrahim, “SQL injection is still alive: A Study on SQL injection signature evasion techniques,” in *Proceedings - 2013 International Conference on Informatics and Creative Multimedia, ICICM 2013*, 2013. doi: 10.1109/ICICM.2013.52.

[10] S. K. Selvaraj, A. Raj, R. Rishikesh Mahadevan, U. Chadha, and V. Paramasivam, “A Review on Machine Learning Models in Injection Molding Machines,” 2022. doi: 10.1155/2022/1949061.

[11] D. Chen, Q. Yan, C. Wu, and J. Zhao, “SQL Injection Attack Detection and Prevention Techniques Using Deep Learning,” in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1757/1/012055.

[12] A. Jain *et al.*, “Overview and Importance of Data Quality for Machine Learning Tasks,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. doi: 10.1145/3394486.3406477.

[13] J. Dean, D. Patterson, and C. Young, “A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution,” *IEEE Micro*, vol. 38, no. 2, 2018, doi: 10.1109/MM.2018.112130030.

[14] M. Takamoto *et al.*, “PDEBENCH: An Extensive Benchmark for Scientific Machine Learning,” in *Advances in Neural Information Processing Systems*, 2022.

[15] R. D․ L․, A. H․ V․, P. B. Honnavalli, and N. S․, “The MongoDB injection dataset: A comprehensive collection of MongoDB - NoSQL injection attempts and vulnerabilities,” *Data Brief*, vol. 54, 2024, doi: 10.1016/j.dib.2024.110289.

[16] N. W. Paton and O. Díaz, “Active Database Systems,” *ACM Comput Surv*, vol. 31, no. 1, 1999, doi: 10.1145/311531.311623.

[17] M. Miftakul Amin, A. Maseleno, K. Shankar, E. Perumal, R. M. Vidhyavathi, and S. K. Lakshmanaprabu, “Active database system approach and rule based in the development of academic information system,” *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 2.26 Special Issue  26, 2018, doi: 10.14419/ijet.v7i2.26.14361.

[18] PortSwigger, “What is cross-site scripting (XSS) and how to prevent it? | Web Security Academy,” portswigger.net.

[19] M. Bravenboer, E. Dolstra, and E. Visser, “Preventing injection attacks with syntax embeddings,” *Sci Comput Program*, vol. 75, no. 7, 2010, doi: 10.1016/j.scico.2009.05.004.

[20] E. F. Codd, “Relational Database: A Practical Foundation for Productivity,” *Commun ACM*, vol. 25, no. 2, 1982, doi: 10.1145/358396.358400.

[21] W. Zhou *et al.*, “Attack sample generation algorithm based on data association group by GAN in industrial control dataset,” *Comput Commun*, vol. 173, 2021, doi: 10.1016/j.comcom.2021.04.014.

[22] Internet World Stats, “World Internet Users Statistics. World Population Stats,” Usage and Population Statistics.

[23] S. N. F. N. B. Mustaffa and M. Farhan, “Detection of False Data Injection Attack using Machine Learning approach,” *Mesopotamian Journal of CyberSecurity*, vol. 2022, 2022, doi: 10.58496/MJCS/2022/005.

[24] S. Lockey, N. Gillespie, D. Holm, and I. A. Someh, “A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021. doi: 10.24251/hicss.2021.664.

[25] Y. Himeur, S. S. Sohail, F. Bensaali, A. Amira, and M. Alazab, “Latest trends of security and privacy in recommender systems: A comprehensive review and future perspectives,” 2022. doi: 10.1016/j.cose.2022.102746.

[26] D. Van Landuyt, V. Wijshoff, and W. Joosen, “A study of NoSQL query injection in Neo4j,” *Comput Secur*, vol. 137, 2024, doi: 10.1016/j.cose.2023.103590.

[27] D. Das, U. Sharma, and D. K. Bhattacharyya, “Defeating SQL injection attack in authentication security: an experimental study,” *Int J Inf Secur*, vol. 18, no. 1, 2019, doi: 10.1007/s10207-017-0393-x.

[28] M. Haššo, O. Sarakhman, S. Đurđić, D. Stanković, and Ľ. Švorc, “Advanced electrochemical platform for simple and rapid quantification of tannic acid in beverages using batch injection analysis with amperometric detection,” *Journal of Electroanalytical Chemistry*, vol. 942, 2023, doi: 10.1016/j.jelechem.2023.117578.

[29] W. Khan, T. Kumar, C. Zhang, K. Raj, A. M. Roy, and B. Luo, “SQL and NoSQL Database Software Architecture Performance Analysis and Assessments—A Systematic Literature Review,” 2023. doi: 10.3390/bdcc7020097.

[30] J. K. Chen and W. Z. Lee, “An introduction of NoSQL databases based on their categories and application industries,” *Algorithms*, vol. 12, no. 5, 2019, doi: 10.3390/a12050106.

[31] F. Chen *et al.*, “Research Progress on Soil Security Assessment in Farmlands and Grasslands Based on Bibliometrics over the Last Four Decades,” *Sustainability (Switzerland)*, vol. 16, no. 1, 2024, doi: 10.3390/su16010404.

[32] J. Brownlee, “Basic Concepts in Machine Learning,” 2015.

[33] A. S. Lea and D. S. Jones, “Mind the Gap — Machine Learning, Dataset Shift, and History in the Age of Clinical Algorithms,” *New England Journal of Medicine*, vol. 390, no. 4, 2024, doi: 10.1056/nejmp2311015.

[34] J. Hegde and B. Rokseth, “Applications of machine learning methods for engineering risk assessment – A review,” 2020. doi: 10.1016/j.ssci.2019.09.015.

[35] C. Virmani, T. Choudhary, A. Pillai, and M. Rani, “Applications of machine learning in cyber security,” in *Research Anthology on Machine Learning Techniques, Methods, and Applications*, 2022. doi: 10.4018/978-1-6684-6291-1.ch033.

[36] U. Tank, S. Arirangan, A. R. Paduri, and N. Darapaneni, “A Study Towards Building Content Aware Models in NLP using Genetic Algorithms,” *EAI Endorsed Transactions on AI and Robotics*, vol. 2, 2023, doi: 10.4108/airo.4078.

[37] T. Lin, Y. Wang, X. Liu, and X. Qiu, “A survey of transformers,” *AI Open*, vol. 3, 2022, doi: 10.1016/j.aiopen.2022.10.001.

[38] A. Saka *et al.*, “GPT models in construction industry: Opportunities, limitations, and a use case validation,” 2024. doi: 10.1016/j.dibe.2023.100300.

[39] K. Han *et al.*, “A Survey on Vision Transformer,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, 2023, doi: 10.1109/TPAMI.2022.3152247.

[40] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient Transformers: A Survey,” *ACM Comput Surv*, vol. 55, no. 6, 2022, doi: 10.1145/3530811.

[41] V. O. Jithin and N. Subramanian, “SECURE-D:Framework For Detecting and Preventing Attacks in SQL and NoSQL Databases,” in *10th International Symposium on Digital Forensics and Security, ISDFS 2022*, 2022. doi: 10.1109/ISDFS55398.2022.9800805.

[42] V. Sachdeva and S. Gupta, “Basic NOSQL Injection Analysis and Detection on MongoDB,” in *2018 International Conference on Advanced Computation and Telecommunication, ICACAT 2018*, 2018. doi: 10.1109/ICACAT.2018.8933707.

[43] A. M., O. H., H. M., and A. S., “NoSQL Racket: A Testing Tool for Detecting NoSQL Injection Attacks in Web Applications,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 11, 2017, doi: 10.14569/ijacsa.2017.081178.

[44] S. Priyadharshini and R. Rajmohan, “Analysis on Database Security Model Against NOSQL Injection,” 2017.

[45] L. Song, Y. Li, and N. Lu, “ProfileSR-GAN: A GAN Based Super-Resolution Method for Generating High-Resolution Load Profiles,” *IEEE Trans Smart Grid*, vol. 13, no. 4, 2022, doi: 10.1109/TSG.2022.3158235.

[46] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, “Trustworthy Artificial Intelligence: A Review,” 2023. doi: 10.1145/3491209.

[47] D. Ray and J. Ligatti, “Defining code-injection attacks,” in *Conference Record of the Annual ACM Symposium on Principles of Programming Languages*, 2012. doi: 10.1145/2103656.2103678.

In the Name of God



**دانشگاه صنعتی مالک‌اشتر**

Malek Ashtar University of Technology

Faculty of Electrical and Computer Engineering

M.Sc. Thesis

Title

**Improving the Detection of پایگاه‌داده غیررابطه‌ای Injection Attacks Based on Machine Learning**

By

Amirhossein Ghasemi

Supervised by

[Supervisor's name(s)]

Advised by

[Advisor's name(s)]

**Dec. 23, 2024**

1. Not Only SQL Injection [↑](#footnote-ref-2)
2. Data Set [↑](#footnote-ref-3)
3. Machin Learning [↑](#footnote-ref-4)
4. SQL [↑](#footnote-ref-5)
5. NoSQL [↑](#footnote-ref-6)
6. OWASP: Open Web Application Security Project [↑](#footnote-ref-7)
7. CVE: Common Vulnerabilities and Exposures [↑](#footnote-ref-8)
8. Firewalls [↑](#footnote-ref-9)
9. Signature based [↑](#footnote-ref-10)
10. Test [↑](#footnote-ref-11)
11. Training [↑](#footnote-ref-12)
12. Natural language processing [↑](#footnote-ref-13)
13. ZeroDay attack is a type of cyber attack exploits a previously unknown security vulnerability for which no patch or solution. [↑](#footnote-ref-14)
14. ImageNet: The project has been instrumental in advancing computer vision and deep learning research. [↑](#footnote-ref-15)
15. TPU: Tensor Processing Units [↑](#footnote-ref-16)
16. Moores law [↑](#footnote-ref-17)
17. Structured [↑](#footnote-ref-18)
18. Semi-Structured [↑](#footnote-ref-19)
19. Non-Structured [↑](#footnote-ref-20)
20. MongoDB is a source-available, cross-platform, document-oriented database program. [↑](#footnote-ref-21)
21. Syntax Injection [↑](#footnote-ref-22)
22. Payload [↑](#footnote-ref-23)
23. Operator Injection [↑](#footnote-ref-24)
24. generative adversarial network, [↑](#footnote-ref-25)
25. BigData [↑](#footnote-ref-26)
26. Hadoop: Apache Hadoop is a collection of open-source software utilities that facilitates using. [↑](#footnote-ref-27)
27. IoT: Internet of Things [↑](#footnote-ref-28)
28. Social technologies [↑](#footnote-ref-29)
29. Cyber Physical Social Systems [↑](#footnote-ref-30)
30. Cybersecurity [↑](#footnote-ref-31)
31. Privacy [↑](#footnote-ref-32)
32. Agile [↑](#footnote-ref-33)
33. JSON: JavaScript Object Notation [↑](#footnote-ref-34)
34. XML: Extensible Markup Language [↑](#footnote-ref-35)
35. Command and Query Responsibility Segregation [↑](#footnote-ref-36)
36. AI [↑](#footnote-ref-37)
37. API: Application Programming Interface [↑](#footnote-ref-38)
38. WAF: Web Application Firewall [↑](#footnote-ref-39)
39. SQLi [↑](#footnote-ref-40)
40. Cloud Computing [↑](#footnote-ref-41)
41. Mr.Das [↑](#footnote-ref-42)
42. dynamic SQL statements [↑](#footnote-ref-43)
43. graph database [↑](#footnote-ref-44)
44. parameterized queries [↑](#footnote-ref-45)
45. Client [↑](#footnote-ref-46)
46. Mr.Young [↑](#footnote-ref-47)
47. Mr.Ramshomar [↑](#footnote-ref-48)
48. Mr. Wen Zhou [↑](#footnote-ref-49)
49. BATADAL: The BATtle of the Attack Detection ALgorithms [↑](#footnote-ref-50)
50. Mr.Dongzhe Lu [↑](#footnote-ref-51)
51. GAN: Generative adversarial network [↑](#footnote-ref-52)
52. Generative [↑](#footnote-ref-53)
53. Discriminators [↑](#footnote-ref-54)
54. latent space [↑](#footnote-ref-55)
55. synthesis-based [↑](#footnote-ref-56)
56. Malicious Payloads [↑](#footnote-ref-57)
57. Legitimate Payloads [↑](#footnote-ref-58)
58. False Positive Rate [↑](#footnote-ref-59)
59. Web scraping [↑](#footnote-ref-60)
60. indexing over array [↑](#footnote-ref-61)
61. embedded objects [↑](#footnote-ref-62)
62. PHP Tautologies Injection [↑](#footnote-ref-63)
63. Javascript Injections [↑](#footnote-ref-64)
64. Piggy-backed Queries [↑](#footnote-ref-65)
65. Escape Sequences [↑](#footnote-ref-66)
66. Carriage Return [CR] [↑](#footnote-ref-67)
67. Line Feed [LF] [↑](#footnote-ref-68)
68. Origin Violation [↑](#footnote-ref-69)
69. Supervised Learning [↑](#footnote-ref-70)
70. Unsupervised Learning [↑](#footnote-ref-71)
71. Reinforcement Learning [↑](#footnote-ref-72)
72. Decession tree [↑](#footnote-ref-73)
73. K-Nearest Neighbors [↑](#footnote-ref-74)
74. Transformer [↑](#footnote-ref-75)
75. Generator [↑](#footnote-ref-76)
76. Discriminator [↑](#footnote-ref-77)
77. Parser Module [↑](#footnote-ref-78)
78. Detection Module [↑](#footnote-ref-79)
79. Log Module [↑](#footnote-ref-80)
80. Parameterized Statements [↑](#footnote-ref-81)
81. Data-Centric Security Model [↑](#footnote-ref-82)
82. Information Gain [↑](#footnote-ref-83)
83. Correlation [↑](#footnote-ref-84)
84. Piggybacked Queries [↑](#footnote-ref-85)
85. 10-fold Cross Validation [↑](#footnote-ref-86)
86. eXtreme Gradient Boosting [↑](#footnote-ref-87)
87. Recall [↑](#footnote-ref-88)
88. Precision [↑](#footnote-ref-89)
89. F1-Score [↑](#footnote-ref-90)
90. ROC-AUC [↑](#footnote-ref-91)
91. Confusion Matrix [↑](#footnote-ref-92)
92. Mean Squared Error [↑](#footnote-ref-93)
93. Mean Absolute Error [↑](#footnote-ref-94)
94. Mean Absolute Percentage Error [↑](#footnote-ref-95)
95. Accuracy [↑](#footnote-ref-96)
96. Area Under the Curve [↑](#footnote-ref-97)
97. Sensitivity [↑](#footnote-ref-98)
98. Specificity [↑](#footnote-ref-99)
99. Sensitivity [↑](#footnote-ref-100)