دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پروژه پنجم هوش مصنوعی

رشته علوم کامپیوتر

تحلیل دیتا

نگارش

علیرضا مختاری

استاد درس

مهدی قطعی

استاد کارگاه

بهنام یوسفی مهر

مهر ۱۴۰۳

چکیده

در این پروژه با استفاده از مجموعه داده‌های مرتبط با حملات XSSو لاگ‌های شبکه، مراحل مختلف تحلیل داده‌ها و اجرای مدل‌های یادگیری ماشین برای شناسایی این حملات انجام شد. ابتدا داده‌ها مورد پیش‌پردازش قرار گرفته و تحلیل اکتشافی داده‌ها (EDA) برای درک الگوها و ویژگی‌های مهم صورت گرفت. سپس از مدل‌های طبقه‌بندی مختلف برای شناسایی حملات XSSاستفاده شد و عملکرد آن‌ها با معیارهایی مانند Accuracyو F1-Scoreارزیابی گردید. در ادامه با بهینه‌سازی مدل‌ها از طریق روش‌هایی مانند Cross Validationو Hyperparameter Tuning، دقت شناسایی بهبود یافت. در نهایت با جلوگیری از مشکلاتی مانند Data Leakageو استفاده از روش‌های تشخیص ناهنجاری Anomaly Detection، مدل نهایی کارایی مطلوبی در شناسایی حملات XSSنشان داد.

واژه‌های کلیدی:

تحلیل داده‌ها، حملات XSS، لاگ شبکه، پیش‌پردازش داده‌ها، تحلیل اکتشافی داده‌ها(EDA) ، طبقه‌بندی، ارزیابی مدل، Cross Validation، Hyperparameter Tuning، تشخیص ناهنجاری (Anomaly Detection)، جلوگیری از .Data Leakage

فهرست مطالب

فصل اول

مقدمه

مقدمه

با گسترش روزافزون تهدیدات امنیتی در فضای وب، شناسایی و جلوگیری از حملات تزریق XSS (Cross-Site Scripting)از اهمیت ویژه‌ای برخوردار است. این حملات به مهاجمان امکان می‌دهد کدهای مخرب را به وب‌سایت‌ها تزریق کرده و از این طریق به اطلاعات کاربران دسترسی پیدا کنند. در این پروژه با استفاده از لاگ‌های شبکه و تحلیل داده‌ها، فرآیند شناسایی حملات XSSبا به‌کارگیری الگوریتم‌های یادگیری ماشین انجام شده است. هدف اصلی این مطالعه، ارزیابی مدل‌های مختلف طبقه‌بندی و ارائه رویکردی بهینه برای شناسایی سریع و دقیق این نوع حملات است. با تمرکز بر پیش‌پردازش داده‌ها، تحلیل اکتشافی، و بهینه‌سازی مدل‌ها، نتایج حاکی از عملکرد مناسب سیستم در شناسایی حملات XSSمی‌باشد .

فصل دوم

مطالعه و خلاصه برداری مقالات مرتبط در حوزه data science

مقالات

۱. تشخیص حملات XSS با استفاده از الگوریتم‌های یادگیری ماشین
در این مقاله، الگوریتم‌های SVMو Random Forestبرای شناسایی حملات XSS مورد استفاده قرار گرفته‌اند .

- **منبع** : Ahmad, I., Hussain, M., & AlGhamdi, A. S. (2019). "Machine Learning Techniques for Web-Based Attacks Detection."
- **لینک مقاله** : https://doi.org/10.1016/j.jksuci.2018.05.003

۲. تحلیل و ارزیابی امنیت سایبری با روش‌های تشخیص ناهنجاری
مقاله‌ای که به استفاده از Anomaly Detectionبرای شناسایی ناهنجاری‌ها در شبکه‌های داده پرداخته و مدل‌هایی مانند Isolation Forestو One-Class SVMرا بررسی کرده است .

- **منبع** : Chandola, V., Banerjee, A., & Kumar, V. (2009). "Anomaly Detection: A Survey." ACM Computing Surveys.
- **لینک مقاله** : https://doi.org/10.1145/1459352.1459353

۳. بهینه‌سازی شناسایی حملات با استفاده از تکنیک‌های Cross Validation و Hyperparameter Tuning
این مقاله به ارزیابی مدل‌های یادگیری ماشین با Cross Validationو بهینه‌سازی هایپرپارامترها پرداخته است .

- **منبع** : Bergstra, J., & Bengio, Y. (2012). "Random Search for Hyper-Parameter Optimization." Journal of Machine Learning Research.
- **لینک مقاله** : http://jmlr.org/papers/v13/bergstra12a.html

٤. نقش تحلیل اکتشافی داده‌ها (EDA) در شناسایی حملات امنیتی
این مقاله تأکید ویژه‌ای بر استفاده از EDAو ابزارهای بصری‌سازی مانند Matplotlibو Seabornبرای درک داده‌ها و تشخیص الگوها دارد .

- **منبع :** Wickham, H., & Grolemund, G. (2016). "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data."
- **لینک مقاله :** https://r4ds.had.co.nz

٥. مقایسه مدل‌های Ensemble در تشخیص حملات XSS
در این مقاله روش‌های ترکیبی مانند Baggingو Boostingبررسی شده‌اند و مشخص شد مدل‌هایی مانند XGBoostعملکرد بالاتری دارند .

- **منبع :** Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference.
- **لینک مقاله :** https://doi.org/10.1145/2939672.2939785

جمع‌بندی نهایی:

منابع و مقالات فوق به‌طور مستقیم به تحلیل داده‌ها، استفاده از الگوریتم‌های یادگیری ماشین، بهینه‌سازی مدل‌ها، و تشخیص ناهنجاری در حملات XSS مرتبط هستند. در این پروژه از رویکردهای مشابه برای شناسایی و تشخیص حملات **XSS**در لاگ‌های شبکه استفاده شده است.

فصل سوم

روش پیشنهادی و نتایج

روش انجام شده در پروژه و نتایج

پیش‌پردازش داده‌ها

- برای آماده‌سازی داده‌های خام جهت استفاده در مدل‌های یادگیری ماشین، فرآیندهای زیر انجام شد:
- **حذف و انتخاب ویژگی‌ها :**با بررسی داده‌ها، ستون‌هایی مانند src_ip، src_port، dst_ip، host، url، user_agent، response_sizeو status_codeبه‌عنوان ویژگی‌های ورودی باقی ماندند.
- **تحلیل همبستگی :**همبستگی ویژگی‌ها با هدف شناسایی ارتباط بین آن‌ها و برچسب attack انجام شد. نتایج نشان داد :
  - status_codeو user_agentبیشترین همبستگی مثبت با برچسب حمله را دارند.
  - سایر ویژگی‌ها مانند src_ipو dst_ipنیز همبستگی‌های کمتری دارند.
- **کدگذاری متغیرهای دسته‌ای :**از روش‌هایی مانند **Label Encoding**برای کدگذاری مقادیر متنی (مانند hostو urlاستفاده شد.
- **مقیاس‌بندی داده‌های عددی :**داده‌ها با استفاده از تکنیک‌های **Min-Max Scaling** مقیاس‌بندی شدند تا تأثیر مقیاس‌های نامتوازن کاهش یابد.

انتخاب و مقایسه مدل‌های یادگیری ماشین

- با استفاده از کتابخانه **Lazy Predict**، مدل‌های مختلف یادگیری ماشین به‌سرعت ارزیابی شدند. نتایج به‌دست‌آمده نشان داد که مدل‌های زیر عملکرد بهتری دارند:
- **Random Forest Classifier**
- **LightGBM Classifier**
- **AdaBoost Classifier**

- **Bagging Classifier**
- تمام مدل‌های بالا دقت (Accuracy) و F1-Score برابر با **1.0** داشتند، که نشان‌دهنده عملکرد عالی آن‌ها در تشخیص حملات است.

بهینه‌سازی مدل Random Forest

- با توجه به نتایج اولیه و کارایی بالای **Random Forest**، این مدل انتخاب و بهینه‌سازی شد:
- **Cost Sensitivity:** برای کاهش اثر داده‌های نامتوازن احتمالی، پارامترهای مدل با حساسیت هزینه بهینه شدند.
- **Cross-Validation:** برای ارزیابی پایداری مدل، از اعتبارسنجی متقابل (Cross-Validation) استفاده شد.
- **نتایج نهایی مدل Random Forest:**
  - **Accuracy:** 1.0
  - **F1-Score:** 1.0
  - **Precision** و **Recall:** هر دو برابر با ۱.۰

استفاده از روش‌های Ensemble

- در ادامه برای بهبود نتایج و جلوگیری از **Overfitting**، از ترکیب چند مدل یادگیری ماشین (Ensemble Methods) استفاده شد.
- از روش‌های ترکیبی مانند **Bagging** و **Boosting** استفاده شد.
- نتایج نهایی مدل Ensemble:
  - **Accuracy:** 0.9976
  - **F1-Score:** 0.9977
  - **Precision:** 1.0
  - **Recall:** 0.9954

- این روش با کاهش اندک در دقت و عملکرد، نتایج پایدار و مطمئنی ارائه کرد.

جمع‌بندی روش پیشنهادی

- با توجه به داده‌های موجود و نتایج حاصل‌شده، روش پیشنهادی به این صورت است:

۱. استفاده از **Random Forest Classifier** به‌عنوان مدل پایه با تنظیم حساسیت هزینه.

۲. ترکیب مدل‌های مختلف با روش) **Ensemble** مانند Bagging و (Boosting برای افزایش پایداری.

۳. انجام پیش‌پردازش کامل داده‌ها شامل کدگذاری، تحلیل همبستگی و مقیاس‌بندی.

۴. استفاده از **Cross-Validation** برای ارزیابی مدل و جلوگیری از بیش‌برازش.(Overfitting)

- این رویکرد توانست با دقت و F1-Score بسیار بالا حملات XSS را با موفقیت شناسایی کند.

فصل چهارم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

جمع بندی و نتیجه گیری

روش پیشنهادی مبتنی بر مدل‌های یادگیری ماشین و تکنیک‌های Ensembleتوانست با دقت و عملکرد بسیار بالا حملات XSS را در داده‌های شبکه شناسایی کند. نتایج نشان می‌دهد که ویژگی‌های status_codeو user_agentنقش کلیدی در تشخیص این نوع حملات دارند. استفاده از پیش‌پردازش مناسب داده‌ها و بهینه‌سازی مدل‌ها، تأثیر به‌سزایی در افزایش دقت و جلوگیری از Overfittingداشت.

در نهایت، این روش می‌تواند به‌عنوان یک راهکار مؤثر برای شناسایی تهدیدات امنیتی در سیستم‌های مبتنی بر لاگ شبکه به کار گرفته شود و در سیستم‌های تشخیص نفوذ (IDS) مورد استفاده قرار گیرد.

منابع و مراجع

[1] https://ieeexplore.ieee.org/document/10128470/

[۲] https://link.springer.com/article/XXXXXX

[3] https://scikit-learn.org/

[4] https://github.com/shankarpandala/lazypredict