# Детекција на мрежни напади со примена на надгледувано машинско учење

Томи Николоски, Сара Илиевска

kti1032022@feit.ukim.edu.mk, kti772022@feit.ukim.edu.mk

***Абстракт*---Во денешното дигитално општество, заканите од мрежни напади се сè почести и поразновидни, што наметнува потреба од напредни и интелигентни решенија за нивна детекција. Денес, никој не е поштеден од напади и сите треба да се свесни за ризикот од постоење на истите и да преземат мерки со кои би се заштитиле себе си. Овој проект се фокусира на примена на некои поедноставни техники од надгледувано машинско учење за детекција на мрежни напади.**

***Клучни зборови---Decision Tree, Random Forest, Logistic Regression, точност на моделите, F2 перформансна метрика, балансирање на класи, класификација***

## I. Вовед

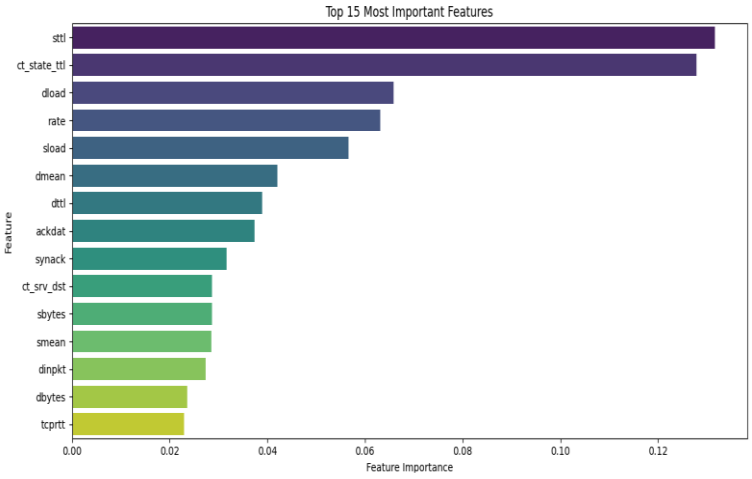
Машинското учење претставува гранка на вештачката интелигенција која се фокусира на развој на алгоритми што учат од податоци и со тек на време го подобруваат своето однесување без експлицитно програмирање. Овие алгоритми можат да откриваат шаблони, да предвидуваат исходи и да носат одлуки врз основа на историски информации. Како едно од најраспространетите и најмоќните алатки на денешницата, машинското учење е широко применето во различни области, вклучително и финансии, автономни возила, и се повеќе во здравството. Исто така, машинското учење станува сè поактуелно и во сајбер безбедноста каде игра сè поголема улога во модерната сајбер безбедност, бидејќи овозможува автоматизирано и интелигентно препознавање на закани. За разлика од традиционалните системи кои се потпираат на однапред дефинирани правила и сигнатури, машинското учење користи историски податоци за да научи шаблони на нормално и сомнително однесување. Ова му овозможува на системот да открие и нови, досега непознати типови на напади кои не би биле препознаени со класични методи, односно не би можеле да се видат со „голо око“. Низ годините луѓето биле сведоции на безброј хакерски напади, од кои некои се врз огромни компании, но со напредокот на технологијата и примената на машинското учење, денес ваквите закани сè почесто се навремено препознаени и спречени, што значително придонесува за поголема сајбер безбедност. Благодарение на алгоритмите за надгледувано машинско учење, денес системите се способни автоматски да идентификуваат сомнително однесување во мрежниот сообраќај. Ова не само што овозможува побрза реакција, туку и значително го намалува бројот на лажно негативни случаи, кои во минатото претставувале голем предизвик за безбедносните аналитичари.

Во овој проект се разработува систем за детекција на мрежни напади користејќи техники на надгледувано машинско учење. Преку анализа на податочниот сет UNSW-NB15[1], кој содржи реалистични симулирани податоци за мрежен сообраќај, се прави предобработка, визуелизација, и примена на различни класификациски алгоритми како Decision Tree, Random Forest и Logistic Regression. Целта е да се изгради модел кој е способен со висока точност да разликува нормален од малициозен сообраќај, како и да се процени перформансот на различни модели преку соодветни метрики. Иако како примарна перформансна метрика се зема точноста на моделот, исто така се зема во предвид и F2 метриката се со цел да се изгради еден стабилен модел кој ќе се обиде да го редуцира бројот на грешки од втор тип, односно лажно да класифицира дека системот не е напднат, а во реалност истиот да биде.

## II. Податочно множество

Во овој проект користено е UNSW-NB15[1] податочно множество креирано од универзитетски истражувачи при Универзитетот на Јужен Велс, во австралискиот град Камбера, кое претставува еден од најпознатите и широко користени множества на податоци во доменот на сајбер безбедноста. Овоа податочно множество е генерирано во реалистично мрежно опкружување и содржи различни типови на мрежен сообраќај, вклучувајќи нормални пакети и пакети поврзани со повеќе видови на напади. Податочното множество се состои од четириесет и девет атрибути, вклучувајќи ја и целната класа, и повеќе од двесте и педесет илјади инстанци, кои опфаќаат различни типови мрежни протоколи, услуги и состојби во мрежата. Целната променлива, класата, која се користи да кажи дали инстанцата која се разгледува во моментот е малициозна или не има вредност 0 за нормален сообраќај и 1 за малициозен напад, што овозможува формулација на задачата како бинарен проблем на класификација. Атрибутите генерално може да се поделат во пет групи и тоа:

- Основни атрубити;
- Статистички атрибути;
- Содржински атрибути;
- Временски-базирани и конекциски-базирани атрибути;
- Целна класа;

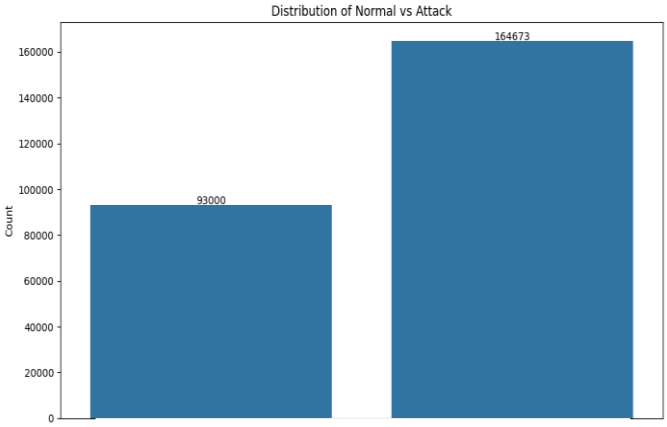



***Слика 1.*** Топ 15 најважни атрибути

Иако извршивме анализа на важноста на карактеристиките со цел да се идентификуваат најважните атрибути, резултатите покажаа дека дури и најважните карактеристики имаат релативно ниска вредност на важност, со максимална вредност од околу 0.13. Ова укажува дека ниедна поединечна карактеристика не доминира значајно во предвидувачката моќ на моделот. Поради ова, беше одлучено да се продолжи со користење на целото множество на атрибути во понатамошната анализа и тренирање на моделите, со цел да се задржи целосната информација содржана во податоците.

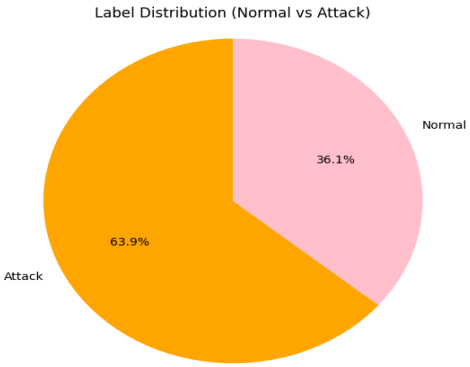
## III. Предпроцесирање и визуелизација

Во делот за предпроцесирање, податочното множество не се промена многу, но сепак претрпе некои промени. Првенствено, во целото множество немаше ниту една испуштена вредност, односно беше целосно пополнето со вредности, што е одлично затоа што не требеше да се занимаваме со решаваање на таков вид на проблем. Се одлучивме да направиме стандардизација на сите нумерички атрибути заради избегнување на можни проблеми околу големите дистанци кои би можеле повеќе да влијаат на моделите, при нивно учење и тренирање, а со стандардизацијата се избегнува таков проблем. Исто така, податочното множество содржеше три категорични атрибути кои ги енкодиравме со помош на техниката лабелирано кодирање, затоа што моделите за машинско учење прават проблем доколку им се дадат вредности кои не се нумерички.




***Слика 2***. Дистрибуција на таргет променливата

На слика 2 е прикажан график кој ја отсликува распространетоста на целната класа, притоа може да се забележи дека прилично доминантна е класата на нападнати примероци во однос на ненападнатите.




***Слика 3.*** Застапеност на таргет променливата

Впрочем, како што можеме да видиме на слика 3, класата на нападнати надвладува со 63.9% наспроти 31.1% на класата на ненападнати. Небалансираноста на калсите често знае да биде голем проблем за моделите на машинско учење, па со цел да избегниме потенцијални резултати кои не би биле релевантни ние направивме техника која се нарекува „undersampling“, односно кратње на примероци од класата која била нападната се со цел да се дојде до некоја изедначеност, при која моделот не би фаворизирал ниту една од класите туку би бил фер према двете. Оваа техника е прилично популарна, а служи за балансираност на класите. На ваков чекор може да се одлучиме затоа што множеството е прилично големо, па мало кратење на истото не би претставувало голем проблем.

Откако го направивме сето ова, следно на ред беше тренирањето на моделите. Ова податочно множество е супер затоа што авторот уште од старт направил два посебни документа во кои имаш различни податоци, на едните да трениараш модели, а на другите да предвидуваш и да ги тестираш моделите, мерејќи им ја точноста, прецизноста, чувствителноста или некоја од другите перформансни метрики.

## IV. Тренирање на моделите

За целите на класификација, беа тренирани неколку модели на надгледувано машинско учење, вклучувајќи Random Forest, Decision Tree и Logistic Regression. Овие модели беа избрани поради нивната ефикасност, стабилност и разновидност во пристапот кон класификација. Decision Tree моделот овозможува лесна интерпретација и визуелизација на одлуките. Овој модел е многу убав и од причина што не е модел од типот на „црна кутија“, односно е способен да го објасни своето решение. Random Forest користи ансамбл на дрвја за да ја зголеми точноста. Од друга страна пак, Logistic Regression претставува едноставен но моќен линеарен модел кој е корисен за бинарна класификација и служи како референтен модел. Овие модели беа тренирани врз основа на предпроцесираното и балансирано податочно множество.

## V. Евалуација

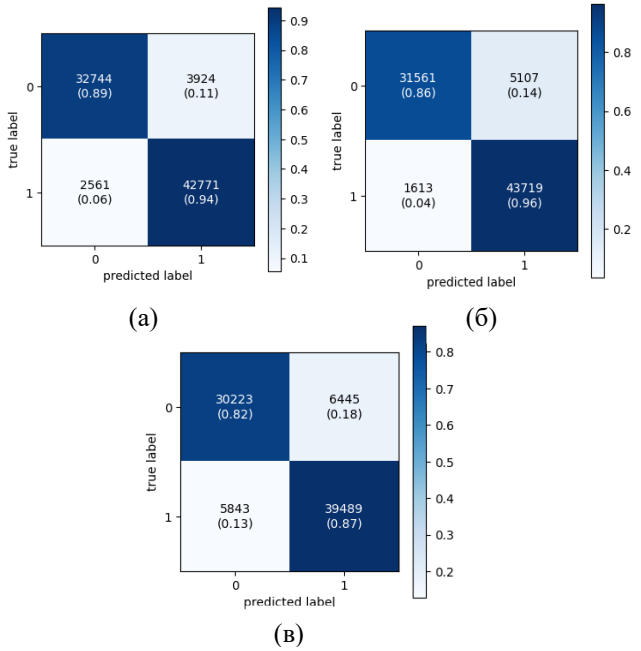
По тренингот на избраните модели, беше извршена евалуација со цел да се процени нивната ефективност и прецизност при класификација на мрежниот сообраќај како нормален или напад. За таа цел беа користени метрики како точност, прецизност, чувствителност, F1 метрика. Овие метрики овозможуваат подетална анализа на перформансите и се индикатор за тоа колку моделите кои се тренирани се добри, односно колку добро научиле и како би се однесувале во реалниот свет.

| Metric | Decision Tree | Random Forest | Logistic Regression |
|---|---|---|---|
| Accuracy | 0.920915 | 0.918049 | 0.850146 |
| Precision | 0.915965 | 0.895404 | 0.859690 |
| Recall | 0.943506 | 0.964418 | 0.871107 |
| F1 Score | 0.929532 | 0.928631 | 0.865361 |

***Табела 1.*** Резултати од моделите

Во табела 1 се прикажани заедно вреднсотите за перформансните метрики на сите три модела. Доколку како приоритет се земе да биде точноста на моделот, тогаш моделот Decision

Tree со 92% е најдобар, но ние покрај точноста како главен индикатор, рековме дека ќе разгледуваме и F2 како перформансна техника. F2 е метрика која ги балансира прецизноста и чувствителноста, придавајќи поголемо значење на чувствителноста, односно способноста на моделот да ги фати сите вистински позитиви. Тоа е особено корисно кога имаат поважност малите грешки во негативните примери, како во случаи на детекција на напади каде промашувањето на напаѓач може да има сериозни последици. F2 е важен кога сакаме да минимизираме грешките од 2 тип. Во продолжение да ги разгледаме матриците на конфузност на моделите




***Слика 4.*** Матрици на конфузност за сите три модела

Може да се забележи дека Random Forest прави убедливо најмалку грешки од 2 тип, 1613, слика 4(б), за разлика од Decision Tree, кој прави 2561, слика 4(а). Односно сето ова изразено нумерички би значело дека Random Forest има F2 вредност од 0.95, додека пак Decision Tree има 0.93. Моделот на логистичка регресија заостанува поприлично зад другите два со F2 еднаква на 0.87. Покрај оваа метрика, логистичката регресија, со 85%, е послаба и со точност во однос на другите два модела.

## VI. Заклучок

Во оваа анализа, разгледавме различни модели на машинско учење за детекција на напади врз компјутерски системи. Применивме неколку класификатори, вклучувајќи Decision Tree, Random Forest и Logistic Regression. По проценката на моделите, се забележа дека Random Forest ја има највисоката вредност за F2 перформансната метрика, што укажува на тоа дека овој модел има подобар баланс помеѓу прецизноста и воспоставување на позитивни случаи во споредба со Decision Tree, кој покажува повисока точност, но помалку добри резултати за критичната метрика како што е F2. Дополнително, преку визуелизација на податоците, беше потврдено дека датасетот е небалансиран, што може да влијае на перформансите на моделите. Токму заради тоа, Random Forest беше посочен како подобар избор, бидејќи има поголема способност за справување со класифицирани податоци, што го прави попогоден за овој вид задача.

Во целост, Random Forest се покажа како најдобар модел за овој проект, со добра способност за справување со класификацијата на нападите врз системите, особено кога се зема предвид балансот помеѓу прецизноста и чувствителноста, што е клучно за оваа задача.

## Референци


[1] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A Comprehensive Dataset for Network Intrusion Detection Systems (IDS). *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS),*

[Online] : https://research.unsw.edu.au/projects/unsw-nb15-dataset

[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

[Online] : RandomForest2001

[3] „Performance Analysis of Intrusion Detection Systems Using a Machine Learning Approach“.

[Online] : Research

[4] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81-106.

[Online] : https://hunch.net/~coms-4771/quinlan.pdf