МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет (национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ АНАЛИЗА СЕТЕВОГО ТРАФИКА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Рецензент:
доцент кафедры ИАОУ, к.т.н.
А.А. Шинкарев

Научный руководитель:
доцент кафедры СП, к.т.н.
М.В. Сухов

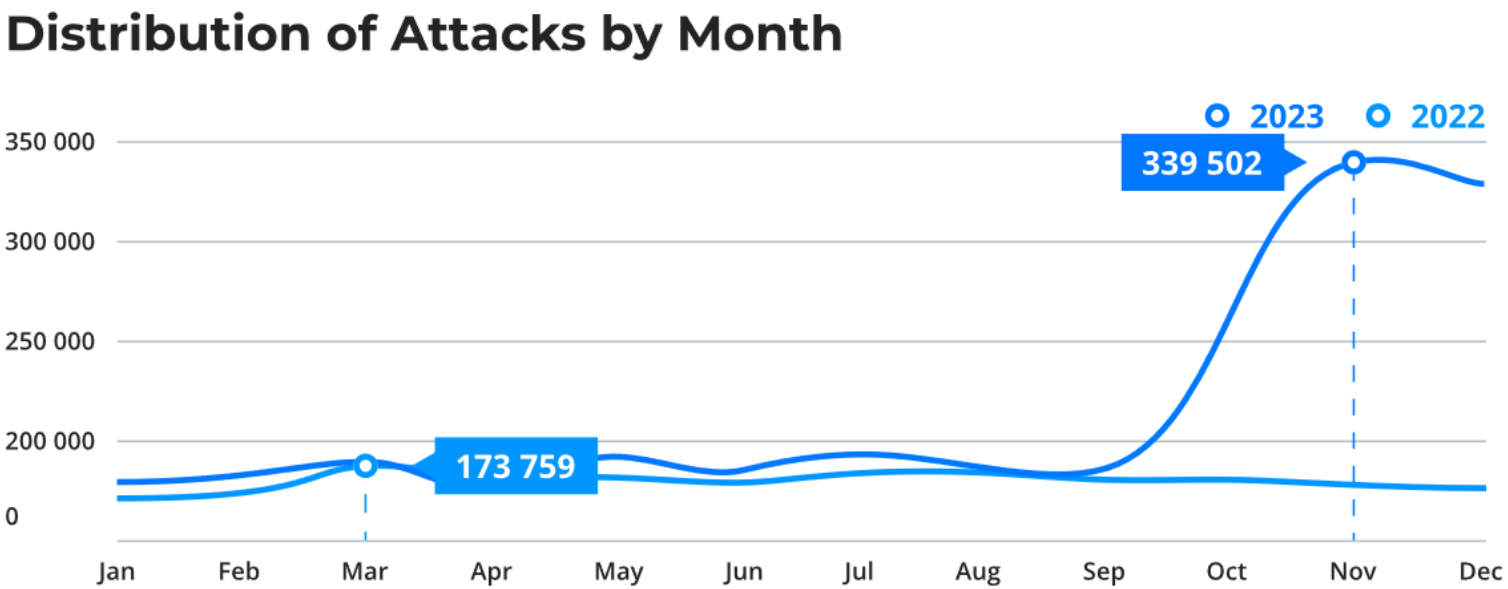
Автор:
студент группы КЭ-228
Г.П. Панюшкин

Челябинск, 2024 г.

АКТУАЛЬНОСТЬ

2 266 198
Total number
of attacks for 2023

1 255 573
Total number
of attacks for 2022



https://ddos-guard.net/en/blog/ddos-attack-trends-2023

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель: Разработка приложения для анализа сетевого трафика в режиме реального времени на основе методов машинного обучения

Задачи:

1. Провести анализ предметной области
2. Собрать набор данных
3. Реализовать и обучить модели машинного обучения
4. Спроектировать систему анализа трафика в реальном времени
5. Реализовать систему анализа интернет-трафика в реальном времени
6. Протестировать систему анализа интернет-трафика в реальном времени

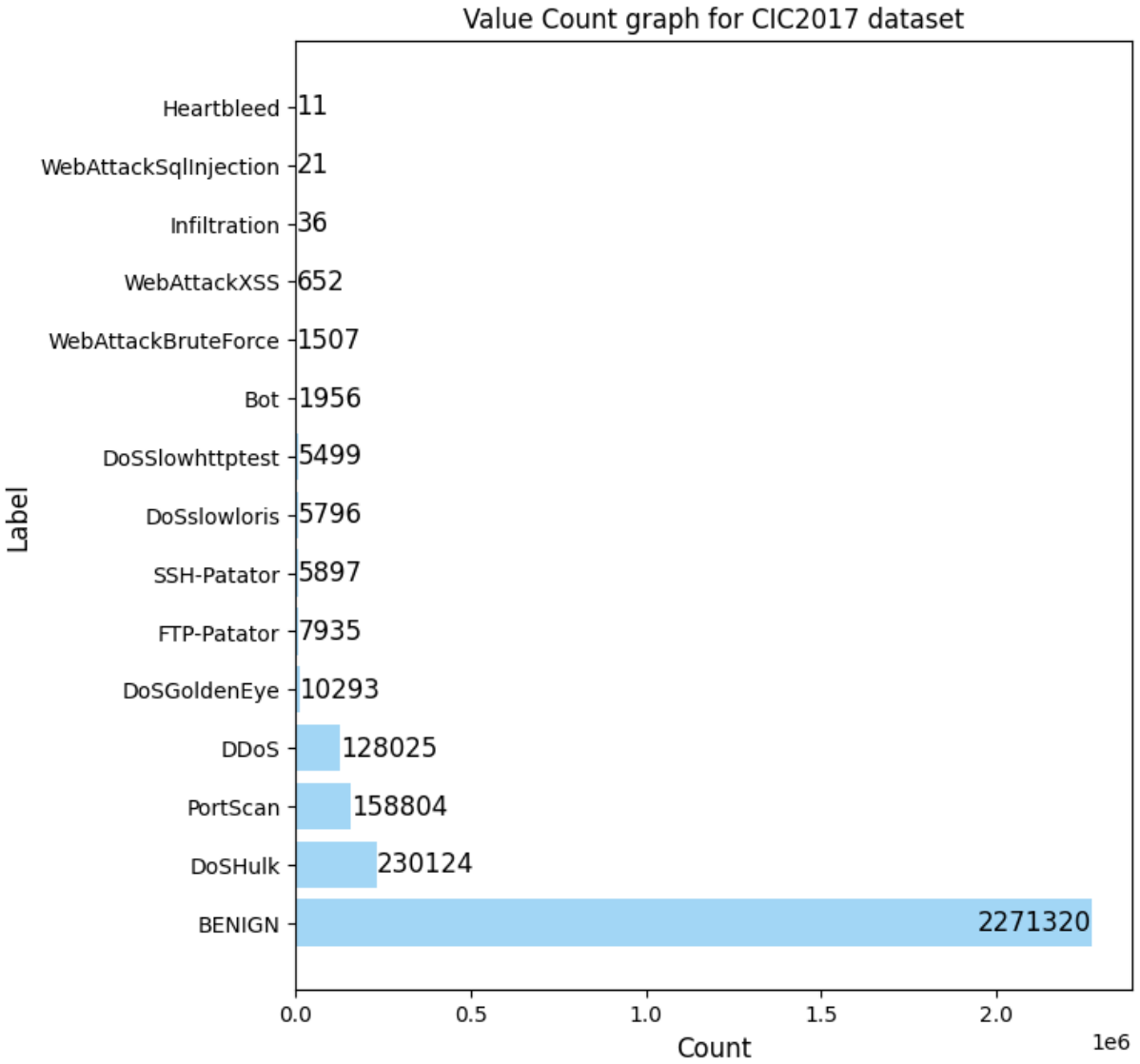
https://github.com/patauch/real-time-DDoS-detection

АНАЛИЗ ЛИТЕРАТУРЫ

Название	Набор данных	Алгоритм	Accuracy
NetGPT: Generative Pretrained Transformer for Network Traffic	ISXW, DoHBrw, USTCTFC, PrivII, Cybermine	NetGPT	98,56%
OMINACS: Online ML-based IOT network attack detection and classification system	Bot-IoT, TON-IOT, CIC-IOT-2022	OMNIACS	98,91%
Comparative analysis of machine learning techniques for network traffic classification	NIMS	K-NN, SVM, Naive Bayes, C4.5 Decision Tree	99,81%
Detecting denial of service attacks using machine learning algorithms	CAIDA 2007	Logistic Regression, Naive Bayes	99,83%

НАБОР ДАННЫХ CIC-IDS2017

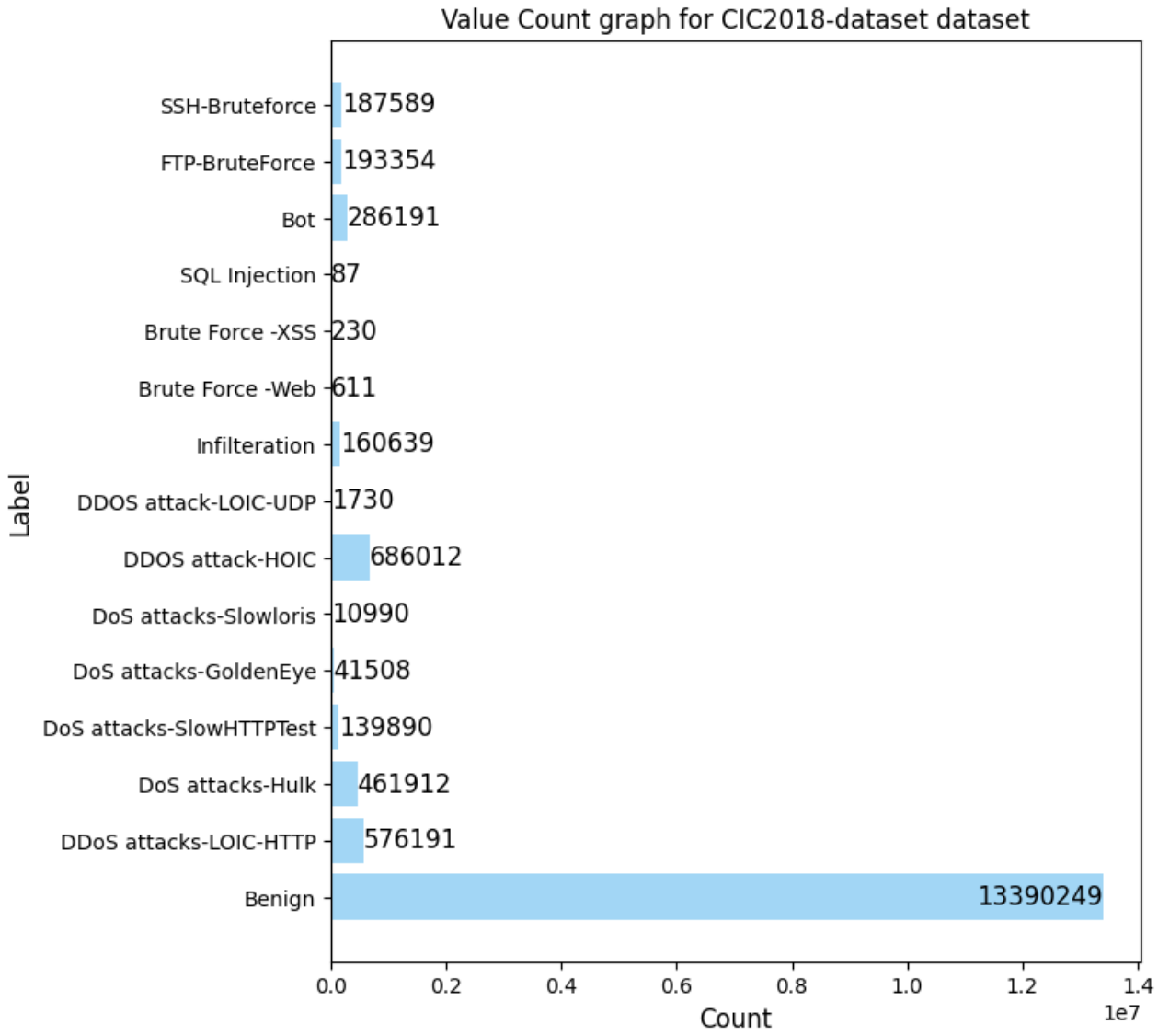
- Количество признаков: 77
- Количество классов: 15
- Количество записей: 2 830 743



https://www.unb.ca/cic/datasets/

НАБОР ДАННЫХ CSE-CIC-IDS2018

- Количество признаков: 80
- Количество классов: 15
- Количество записей: 16 232 943



https://www.unb.ca/cic/datasets/

ПРЕДОБРАБОТКА

1. Удаление пустых записей, дублирующих признаков и признаков с единственным значением
   - Получено 69 признаков
2. Удаление классов с малым количеством записей
   - CIC-IDS2017: 9 классов
   - CSE-CIC-IDS2018: 11 классов
3. Балансировка
   - CIC-IDS2017: 5 499 записей на класс
   - CSE-CIC-IDS2018: 10 990 записей на класс

https://github.com/patauch/real-time-DDoS-detection

ВЫДЕЛЕНИЕ ПРИЗНАКОВ

Получение важности признаков:

• Обученная модель Random Forest

Удаление признаков:

• Recursive Feature Elimination

Итог:

90% информации содержится в 38 признаках из 69

https://github.com/patauch/real-time-DDoS-detection

ИТОГ ПРЕДОБРАБОТКИ

Количество признаков:

• 38 признаков

Количество классов:

• CIC-IDS2017 – 9 классов

• CSE-CIC-IDS2018 – 11 классов

Количество записей на класс:

• CIC-IDS2017 – 5 499

• CSE-CIC-IDS2018 – 10 990

https://github.com/patauch/real-time-DDoS-detection

СРЕДСТВА РАЗРАБОТКИ

Язык программирования: Python 3.12.0

Редактор исходного кода: VSCode

Среда разработки модели машинного обучения:

Google Colaboratory

Библиотеки машинного обучения:

scikit-learn, PyTorch-Lightning, pandas, Matplotlib, NumPy.

Библиотеки приложения

Scapy, pyqt6.

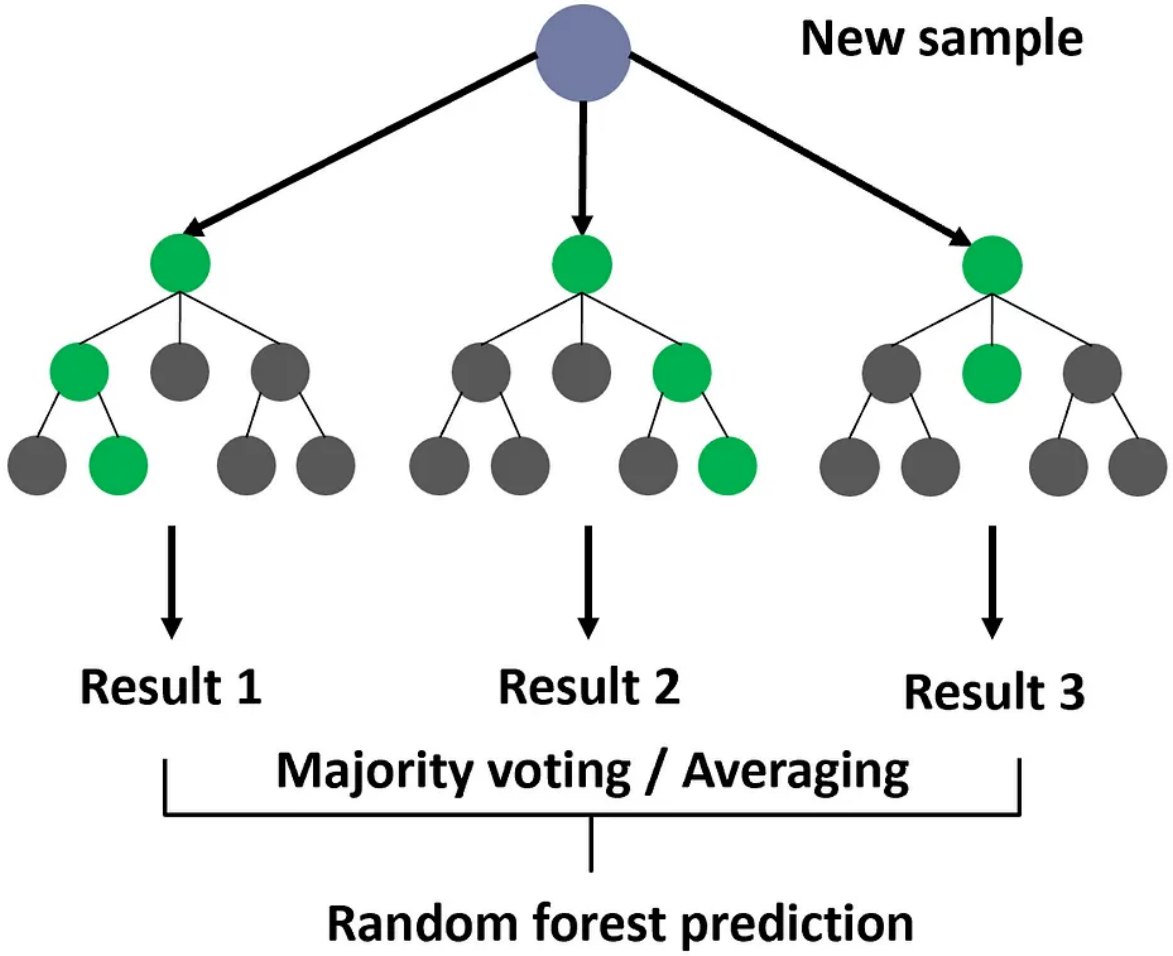
https://github.com/patauch/real-time-DDoS-detection

ИСПОЛЬЗУЕМЫЕ АЛГОРИТМЫ

- Random Forest
- AdaBoost
- CatBoost
- Support Vector Machine
- Long Short Term Memory

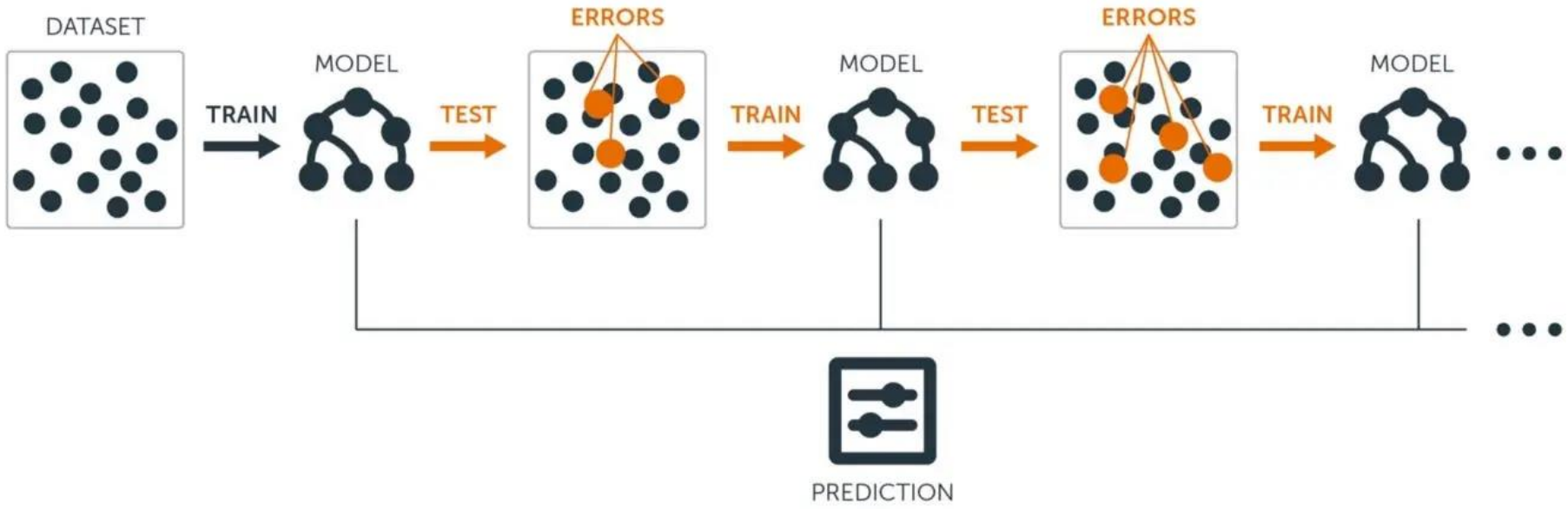
https://github.com/patauch/real-time-DDoS-detection

RANDOM FOREST



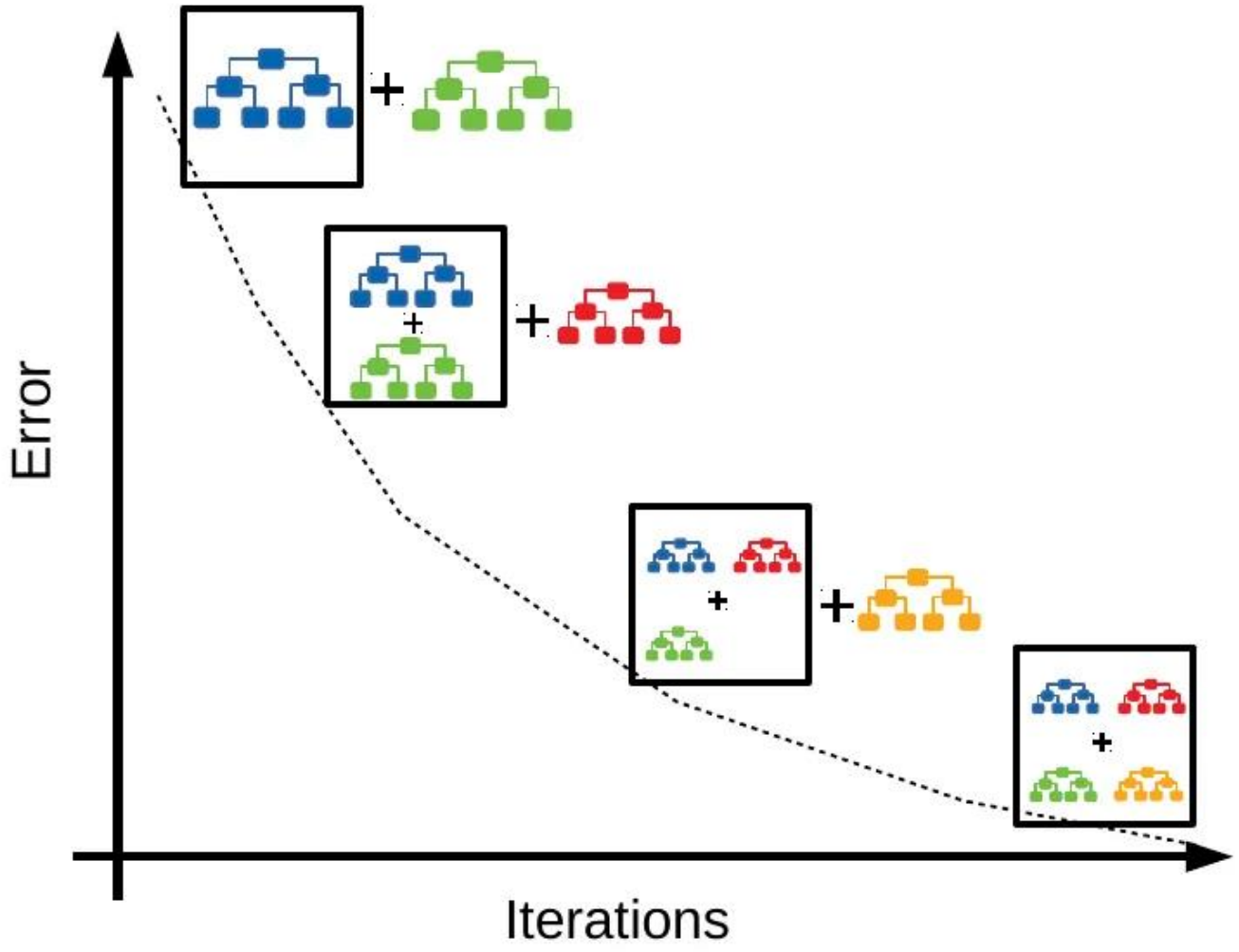
https://medium.com/@roiyeho/random-forests-98892261dc49

ADABOOST



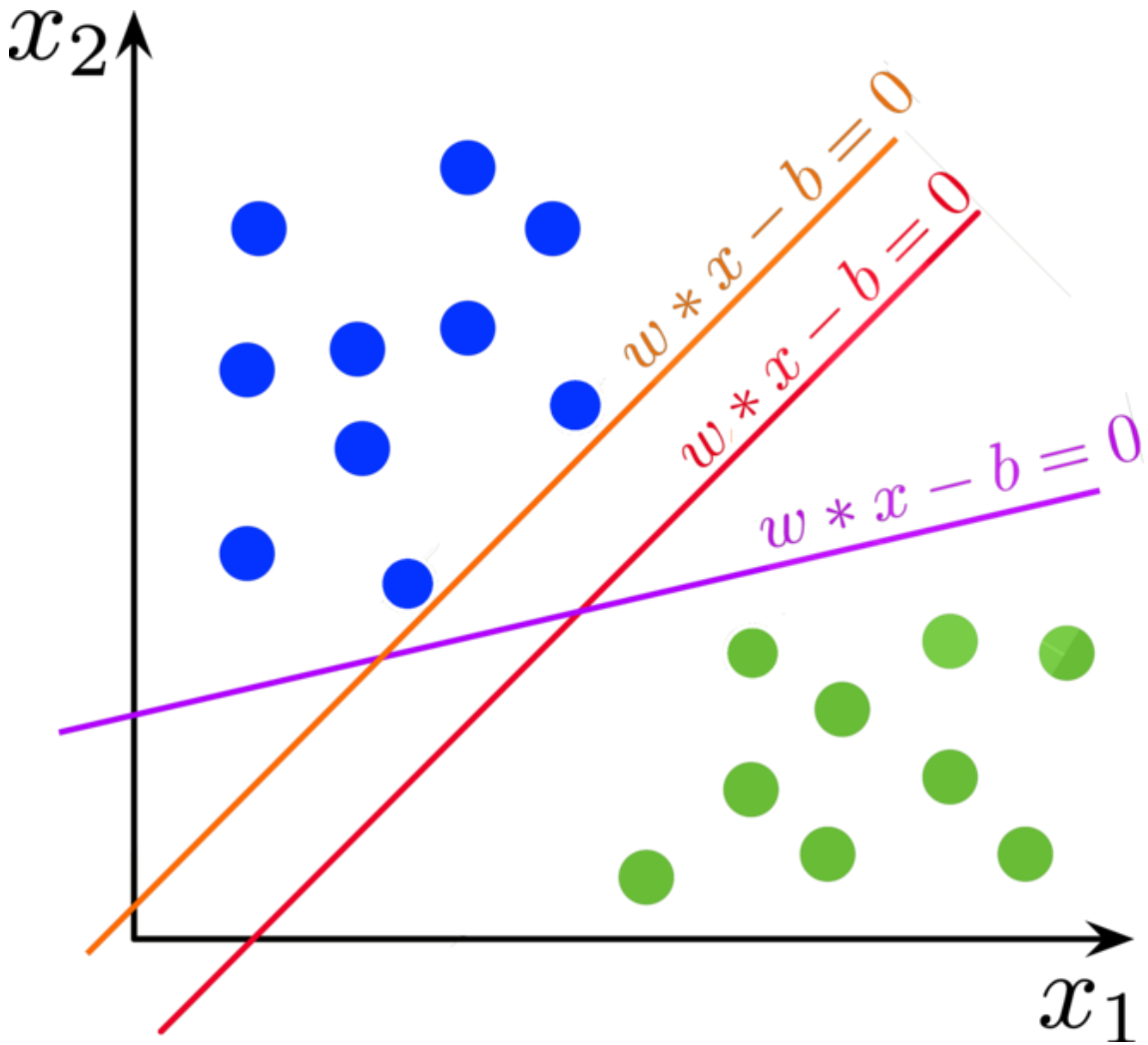
https://scmax.ru/articles/196250/

CATBOOST



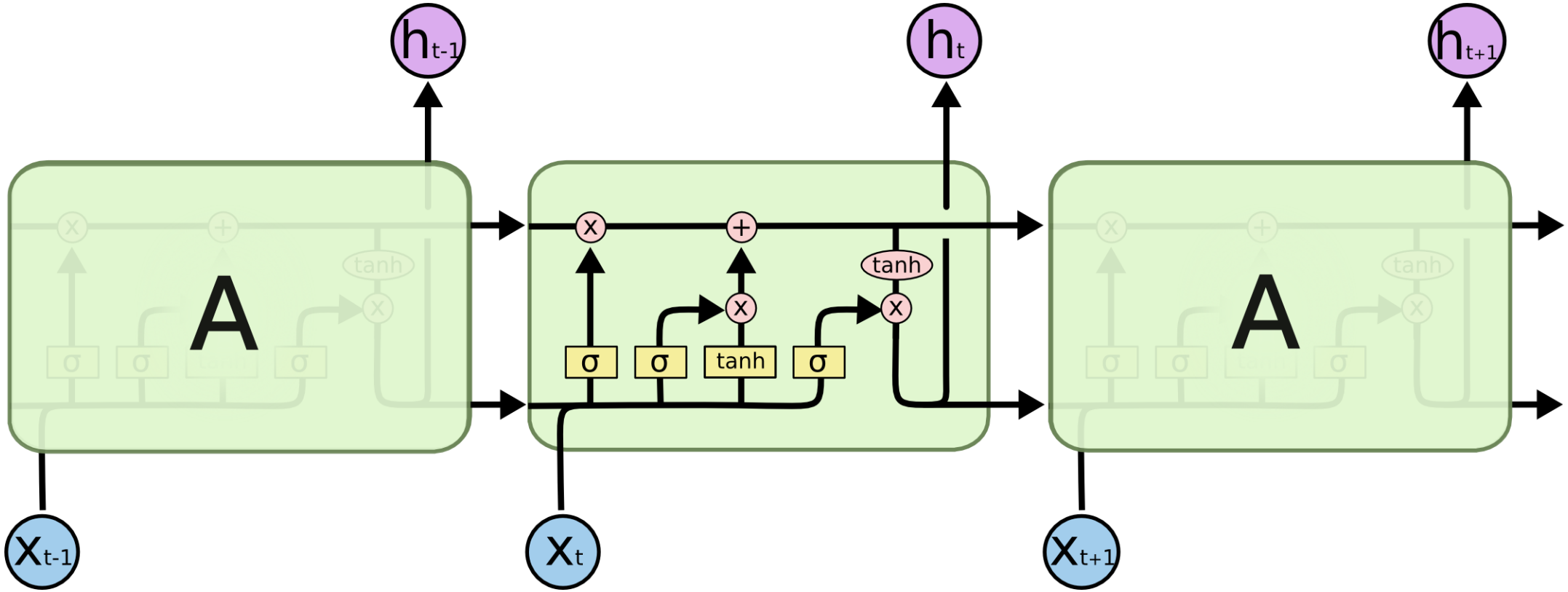
https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

SUPPORT VECTOR MACHINE



https://neerc.ifmo.ru/wiki/index.php?title=Метод_опорных_векторов_(SVM)

LONG SHORT TERM MEMORY



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

ОБУЧЕНИЕ МОДЕЛЕЙ

Обучение (80% набора данных):

- CIC-IDS2017: 39 592 записей из 49 491
- CSE-CIC-IDS2018: 96 712 записей из 120 890

Тестирование (20% набора данных) :

- CIC-IDS2017: 9 899 записей из 49 491
- CSE-CIC-IDS2018: 24 178 записей из 120 890

Платформа:

- Google Colaboratory (GPU: Nvidia Tesla T4, CPU: Intel Xeon CPU @ 2.30GHz)

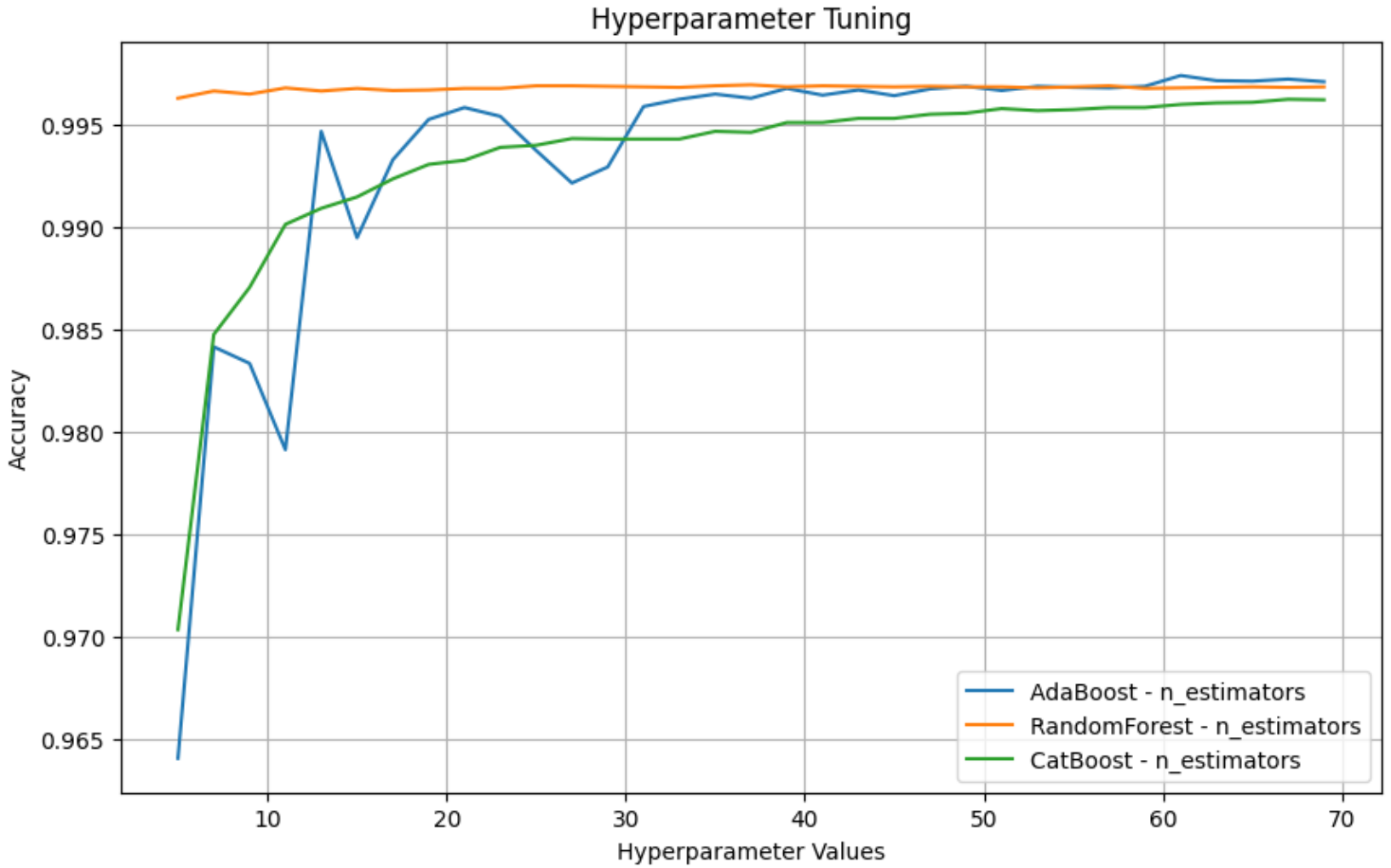
https://github.com/patauch/real-time-DDoS-detection

ОЦЕНКА КАЧЕСТВА ОБУЧЕННЫХ МОДЕЛЕЙ

Набор данных	Random Forest	AdaBoost	SVM	LSTM	CatBoost
CIC-IDS2017	99,8%	99,0%	93,4%	98,3%	99,7%
CSE-CIC-IDS2018	89,4%	89,1%	88,1%	89,3%	90,9%

https://github.com/patauch/real-time-DDoS-detection

ПОДБОР ПАРАМЕТРОВ МОДЕЛИ

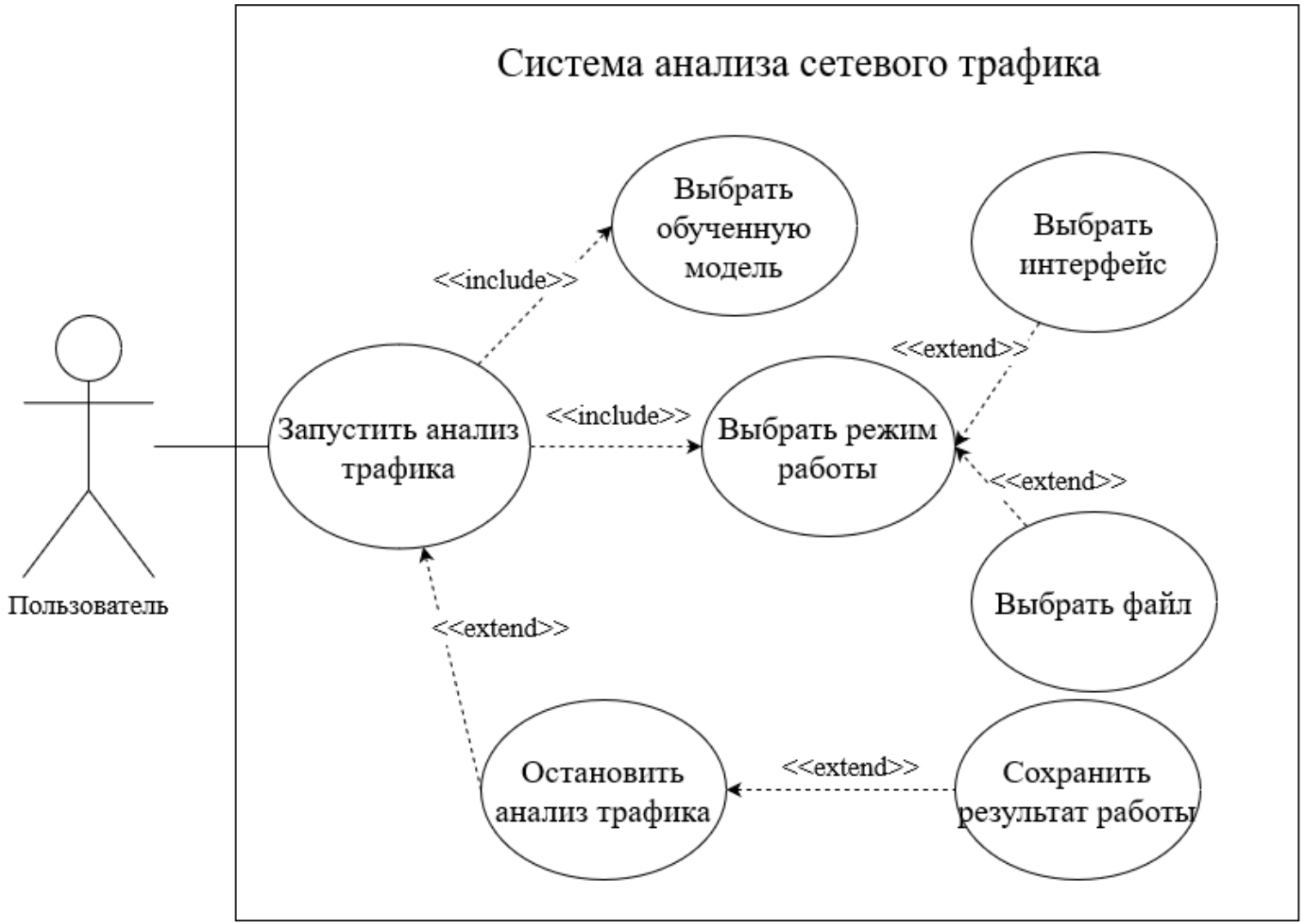


https://github.com/patauch/real-time-DDoS-detection

A/B ТЕСТИРОВАНИЕ

Данные	Accuracy				
	Random Forest	AdaBoost	SVM	CatBoost	LSTM
9 классов, нет балансировки	99,8%	98,9%	85,3%	99,8%	98,1%
9 классов, 1 000 записей на класс	99,3%	98,7%	92,3%	99,5%	94,1%
9 классов, 2 000 записи на класс	99,6%	94,0%	94,5%	99,8%	97,3%
9 классов, 3 000 записи на класс	99,5%	99,5%	94,7%	99,7%	97,4%
9 классов, 5 499 записи на класс	**99,8%**	**99,3%**	**95,4%**	**99,8%**	**98,5%**

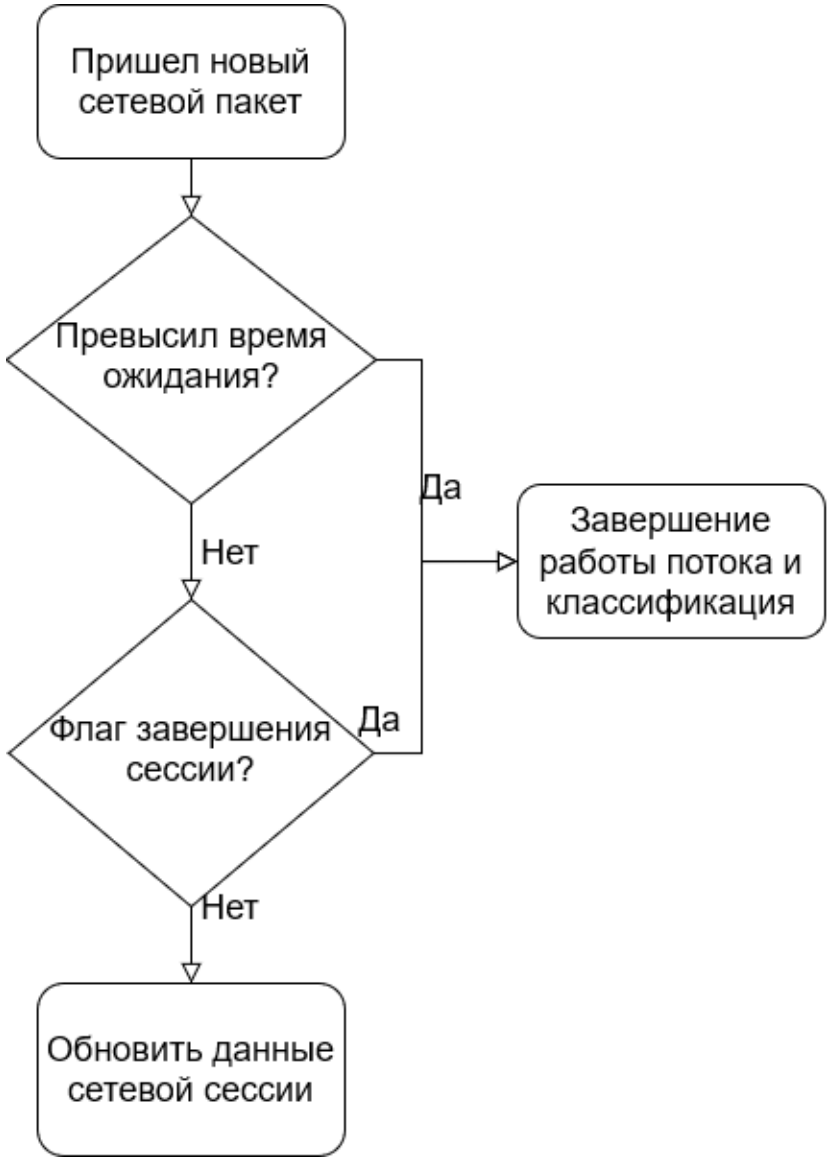
ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ



https://github.com/patauch/real-time-DDoS-detection

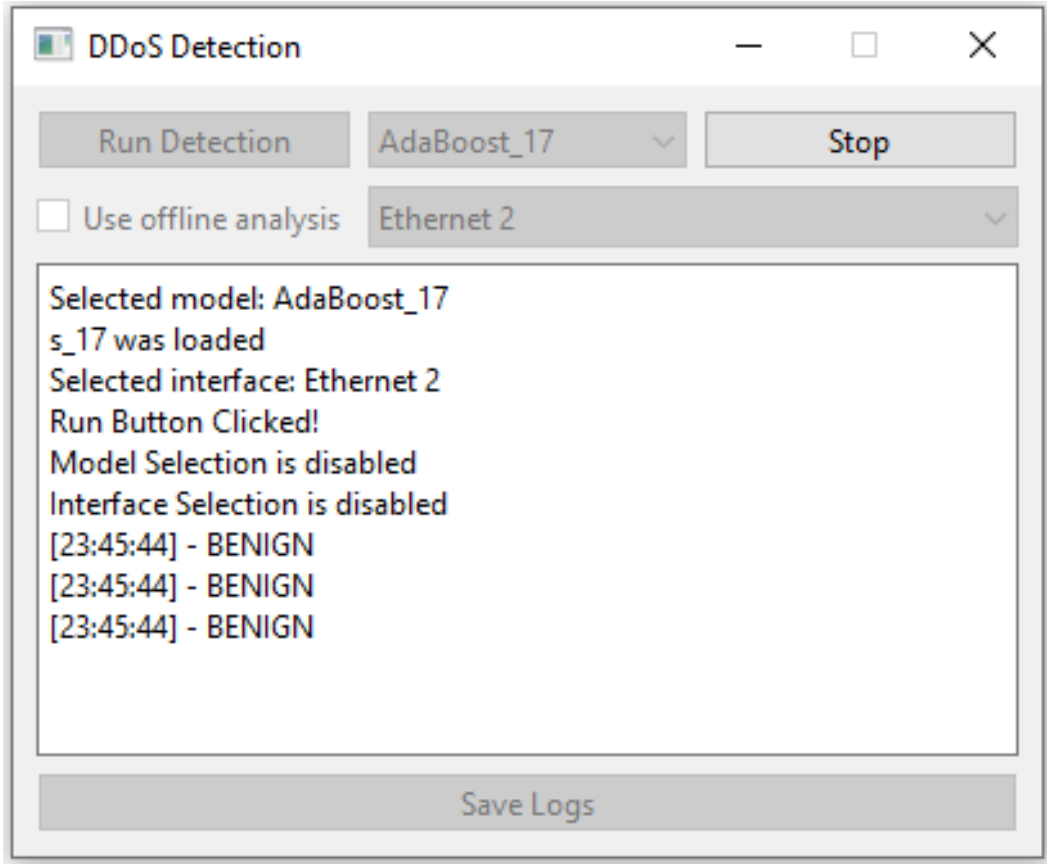
АНАЛИЗАТОР ПАКЕТОВ

- AsyncSniffer из библиотеки Scapy
- Получение информации из интернет-пакета и управление сетевыми потоками
- Классификация информации и вывод результатов



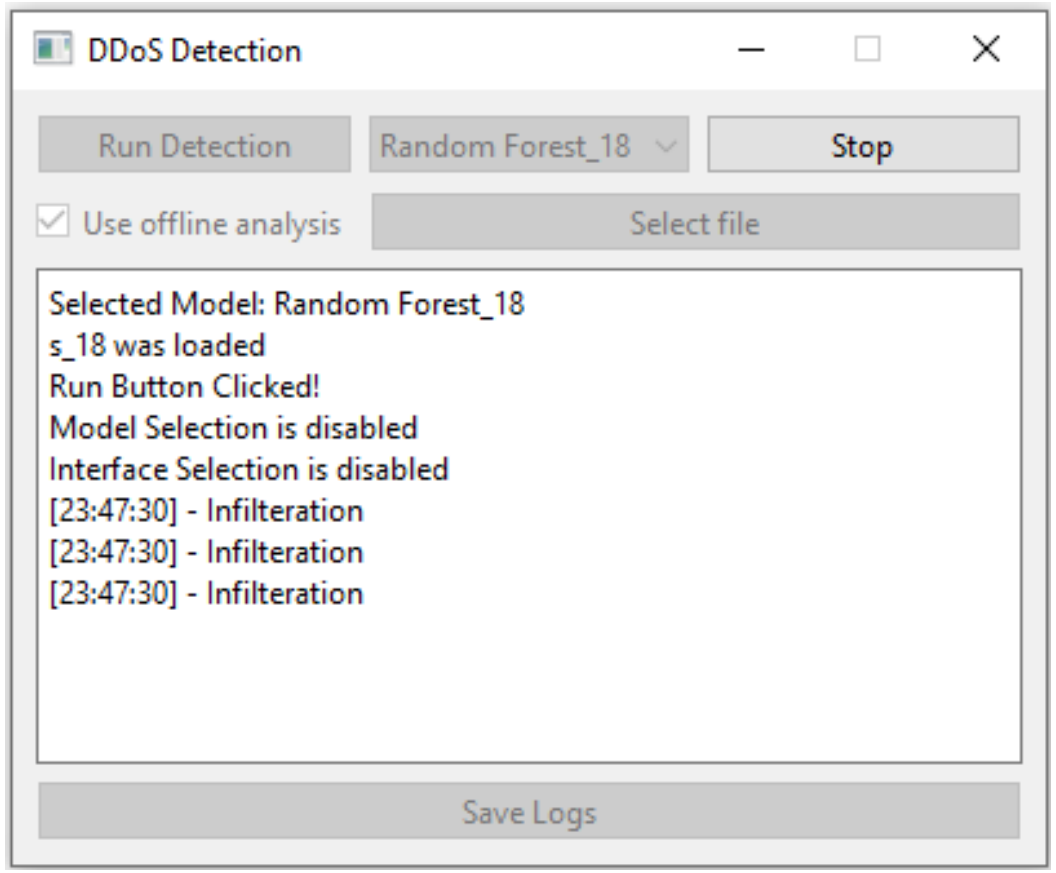
https://github.com/patauch/real-time-DDoS-detection

ТЕСТИРОВАНИЕ РАБОТЫ ПРИЛОЖЕНИЯ НА РЕАЛЬНЫХ ДАННЫХ



https://github.com/patauch/real-time-DDoS-detection

ТЕСТИРОВАНИЕ РАБОТЫ ПРИЛОЖЕНИЯ НА ВРЕДОНОСНОМ ТРАФИКЕ



https://ordo.open.ac.uk/articles/dataset/HTTP_DoS_Dataset_in_PCAP_format_for_Wireshark/17206289/2

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведен анализ предметной области
2. Собран набор данных
3. Реализованы и обучены модели машинного обучения
4. Спроектирована система анализа трафика в реальном времени
5. Реализована система анализа интернет-трафика в реальном времени
6. Протестирована система анализа интернет-трафика в реальном времени

ВАЖНОСТЬ ПРИЗНАКОВ