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

Применение методов машинного обучения для анализа структуры сетевого трафика

Рецензент:
заместитель директора по информационным технологиям
П.Л. Заостровных

Научный руководитель:
доцент кафедры СП, к.ф.-м.н.
А.Т. Латипова

Автор:
студент группы КЭ-220
В.А. Лисовец

АКТУАЛЬНОСТЬ

- Улучшенная защита сети
- Увеличение DDoS-атак
- Адаптация к новым атакам

https://ddos-guard.net/ru/blog/tendentsii-ddos-atak-2023

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель:

Применение методов машинного обучения для анализа структуры сетевого трафика

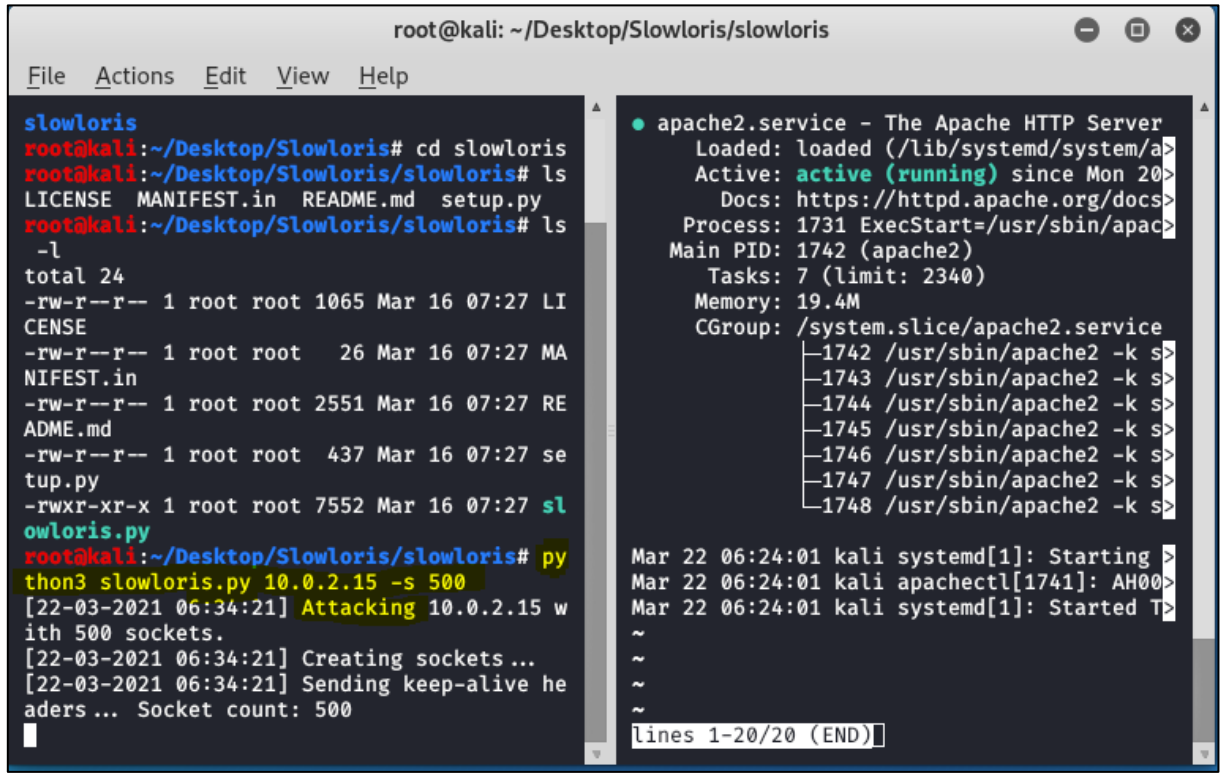
Задачи:

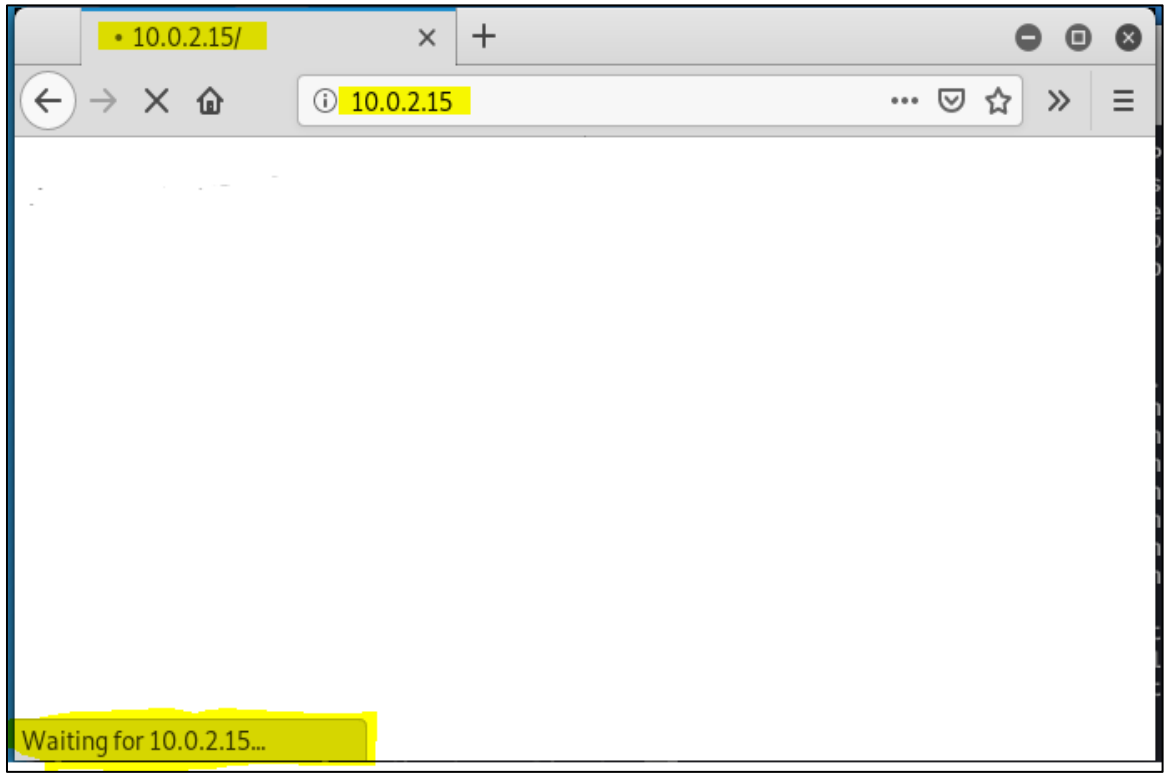
1. Сбор интернет-трафика для тестирования разработанных моделей
2. Разработка, обучение и тестирование алгоритмов машинного обучения
3. Тестирование разработанных алгоритмов на собственном наборе данных

ОБЗОР НАУЧНОЙ ЛИТЕРАТУРЫ

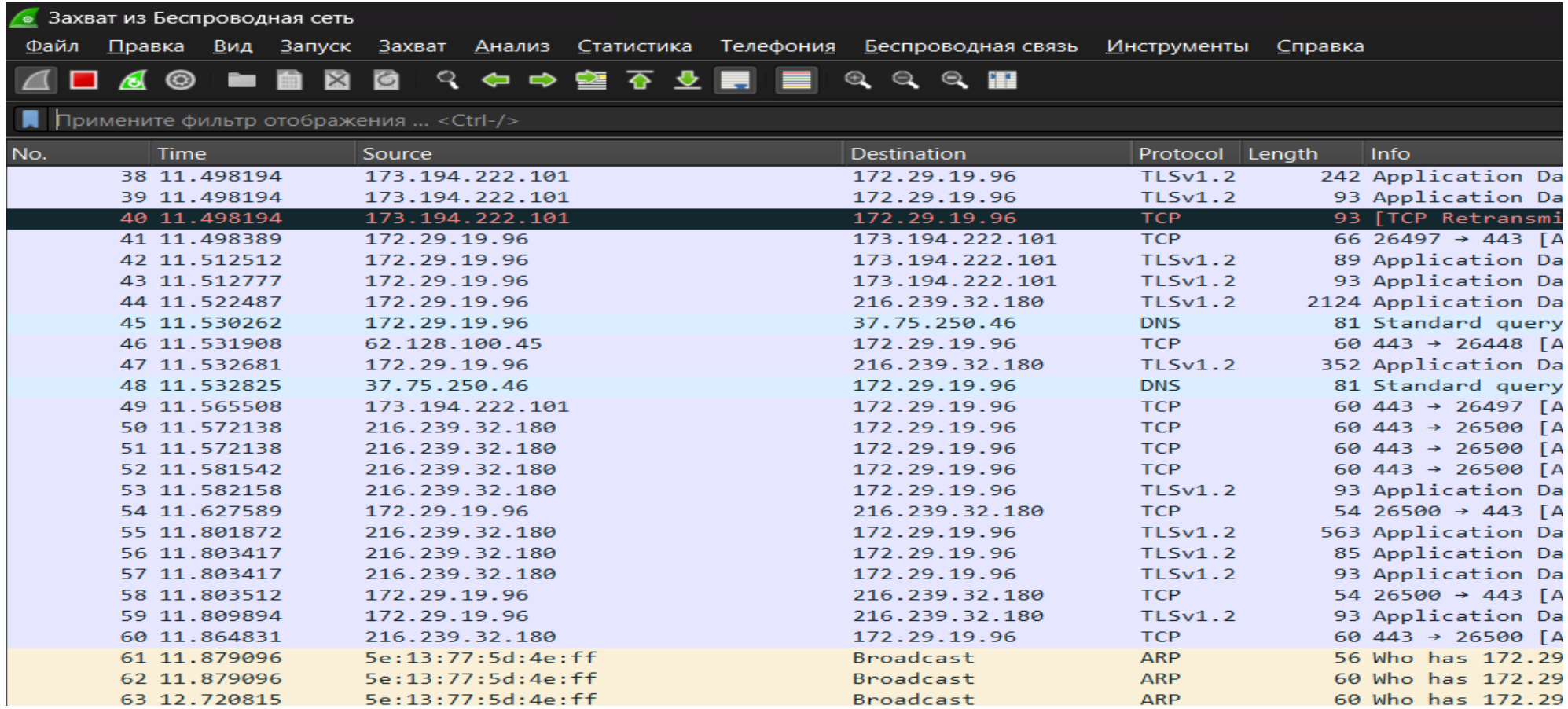
1. Александр К., Евгений Х. Машинное обучение для анализа сетевого трафика. – СПАРК, 2019. – 596 с.
2. Шайлендра С., Гопал К. С. Анализ трафика с использованием алгоритмов машинного обучения. – ICCCA, 2021. – 608 с.
3. Лео Брейман Random Forests. Machine Learning. – UC Berkeley, 2001. – 34 с.
4. Аманжолов, Олжас Маратулы. Исследование методов и средств обнаружения DDoS-атак. – Молодой ученый, 2023. – 5-8 с.

Эмуляция DDoS-атаки





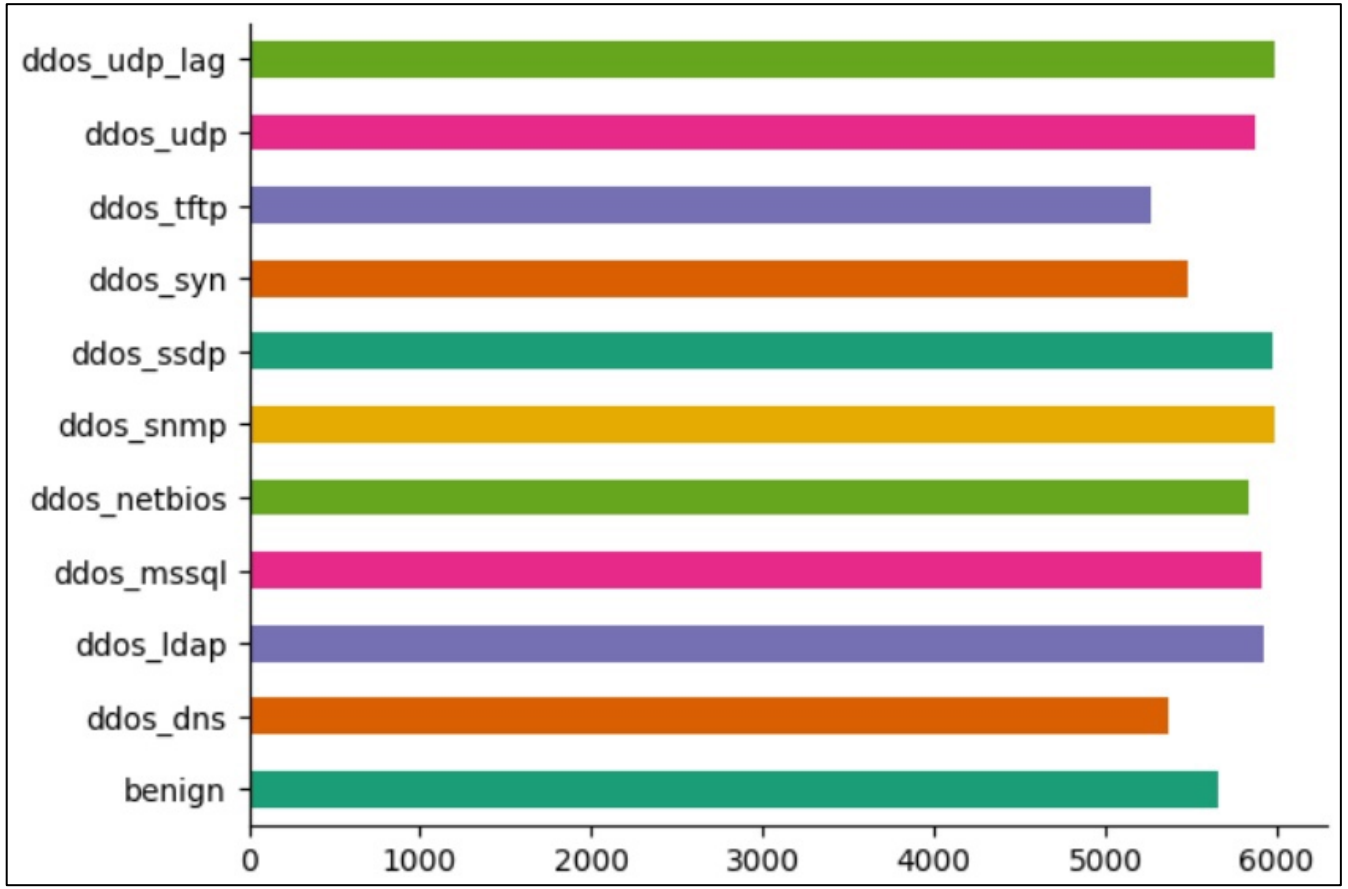
СБОР ТРАФИКА ДЛЯ ТЕСТИРОВАНИЯ



No.	Time	Source	Destination	Protocol	Length	Info
38	11.498194	173.194.222.101	172.29.19.96	TLSv1.2	242	Application Da
39	11.498194	173.194.222.101	172.29.19.96	TLSv1.2	93	Application Da
40	11.498194	173.194.222.101	172.29.19.96	TCP	93	[TCP Retransmi
41	11.498389	172.29.19.96	173.194.222.101	TCP	66	26497 → 443 [A
42	11.512512	172.29.19.96	173.194.222.101	TLSv1.2	89	Application Da
43	11.512777	172.29.19.96	173.194.222.101	TLSv1.2	93	Application Da
44	11.522487	172.29.19.96	216.239.32.180	TLSv1.2	2124	Application Da
45	11.530262	172.29.19.96	37.75.250.46	DNS	81	Standard query
46	11.531908	62.128.100.45	172.29.19.96	TCP	60	443 → 26448 [A
47	11.532681	172.29.19.96	216.239.32.180	TLSv1.2	352	Application Da
48	11.532825	37.75.250.46	172.29.19.96	DNS	81	Standard query
49	11.565508	173.194.222.101	172.29.19.96	TCP	60	443 → 26497 [A
50	11.572138	216.239.32.180	172.29.19.96	TCP	60	443 → 26500 [A
51	11.572138	216.239.32.180	172.29.19.96	TCP	60	443 → 26500 [A
52	11.581542	216.239.32.180	172.29.19.96	TCP	60	443 → 26500 [A
53	11.582158	216.239.32.180	172.29.19.96	TLSv1.2	93	Application Da
54	11.627589	172.29.19.96	216.239.32.180	TCP	54	26500 → 443 [A
55	11.801872	216.239.32.180	172.29.19.96	TLSv1.2	563	Application Da
56	11.803417	216.239.32.180	172.29.19.96	TLSv1.2	85	Application Da
57	11.803417	216.239.32.180	172.29.19.96	TLSv1.2	93	Application Da
58	11.803512	172.29.19.96	216.239.32.180	TCP	54	26500 → 443 [A
59	11.809894	172.29.19.96	216.239.32.180	TLSv1.2	93	Application Da
60	11.864831	216.239.32.180	172.29.19.96	TCP	60	443 → 26500 [A
61	11.879096	5e:13:77:5d:4e:ff	Broadcast	ARP	56	Who has 172.29
62	11.879096	5e:13:77:5d:4e:ff	Broadcast	ARP	60	Who has 172.29
63	12.720815	5e:13:77:5d:4e:ff	Broadcast	ARP	60	Who has 172.29

НАБОР ДАННЫХ ДЛЯ ОБУЧЕНИЯ

CSV-файл для обучения взят на сайте kaggle.com
(https://www.kaggle.com/datasets/devendra416/ddos-datasets)
Общее количество признаков набора данных: 20
Количество записей в наборе данных: 64 249



ПРЕДОБРАБОТКА НАБОРА ДАННЫХ

1. Чистка и балансировка данных
2. Преобразование категориальных данных в числовые
3. Отбор признаков по матрице корреляции

Итог:

- Из 20 признаков осталось 11
- Записей вместо 64 249 стало 64 239
- Все данные преобразованы в числовые

Метрика машинного обучения.

- Recall

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Precision

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- F1-score

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

СРЕДСТВА РАЗРАБОТКИ

Язык программирования: Python 3.10.12

Редактор исходного кода: Google Colab

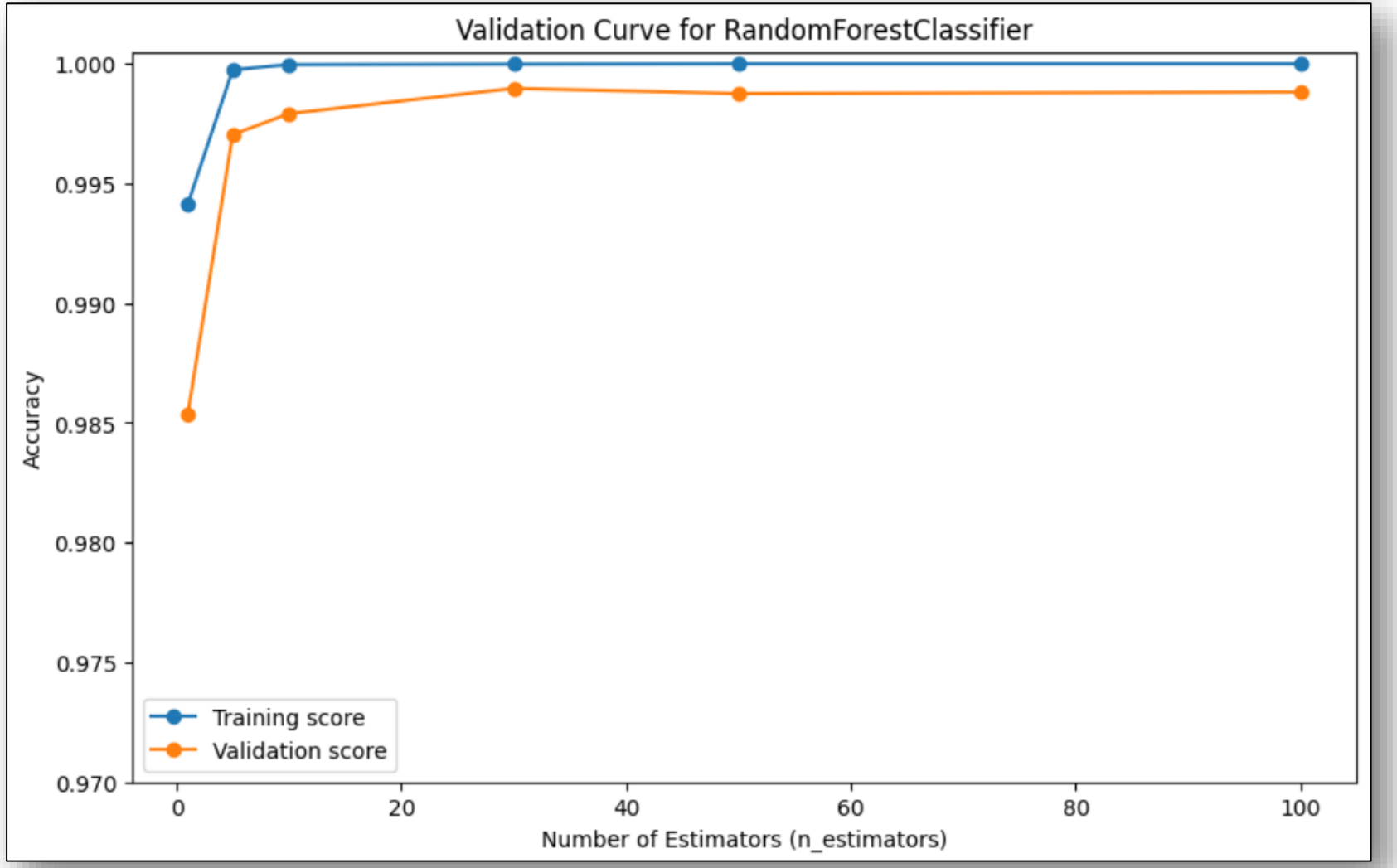
https://colab.research.google.com/drive/1xbBtEXstV79KwrA9fpT9UL0XLTsVCBmK?usp=sharing

Библиотеки для машинного обучения: pandas 2.0.3, scikit-learn 1.2.2, numpy 1.25.2, matplotlib 3.8.3

НАБОР ПАРАМЕТРОВ МОДЕЛЕЙ

Модель	Набор параметров
K-NN	n_neighbors =1, 3, 5, 7, 9
RandomForest	param_range = 1, 5, 10, 30, 50, 100
Logistic Regression	max_iter=10000, solver='saga'

ОБУЧЕНИЕ МОДЕЛЕЙ



РЕЗУЛЬТАТ ОБУЧЕНИЯ МОДЕЛЕЙ

Модель	Время обучения	Метрика	Значение
K-NN	3 минуты 11 секунд	accuracy (точность)	98.13%
RandomForest	34 секунды	accuracy (точность)	99.52%
Logistic Regression	2 часа 03 минуты	accuracy (точность)	98.65%

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Собран интернет-трафик для тестирования разработанных моделей
2. Разработаны, обучены и протестированы алгоритмы машинного обучения

ДАЛЬНЕЙШАЯ РАБОТА

- Добавление большего количества моделей
- Создание веб-приложения