ФГБОУ ВО Нижегородский государственный технический университет им. Р.Е. Алексеева

Выпускная квалификационная работа

Модель и алгоритмы идентификации пользователя по сетевому трафику

Выполнила: студентка гр. М18-ИВТ-2 Ефоде И.М.

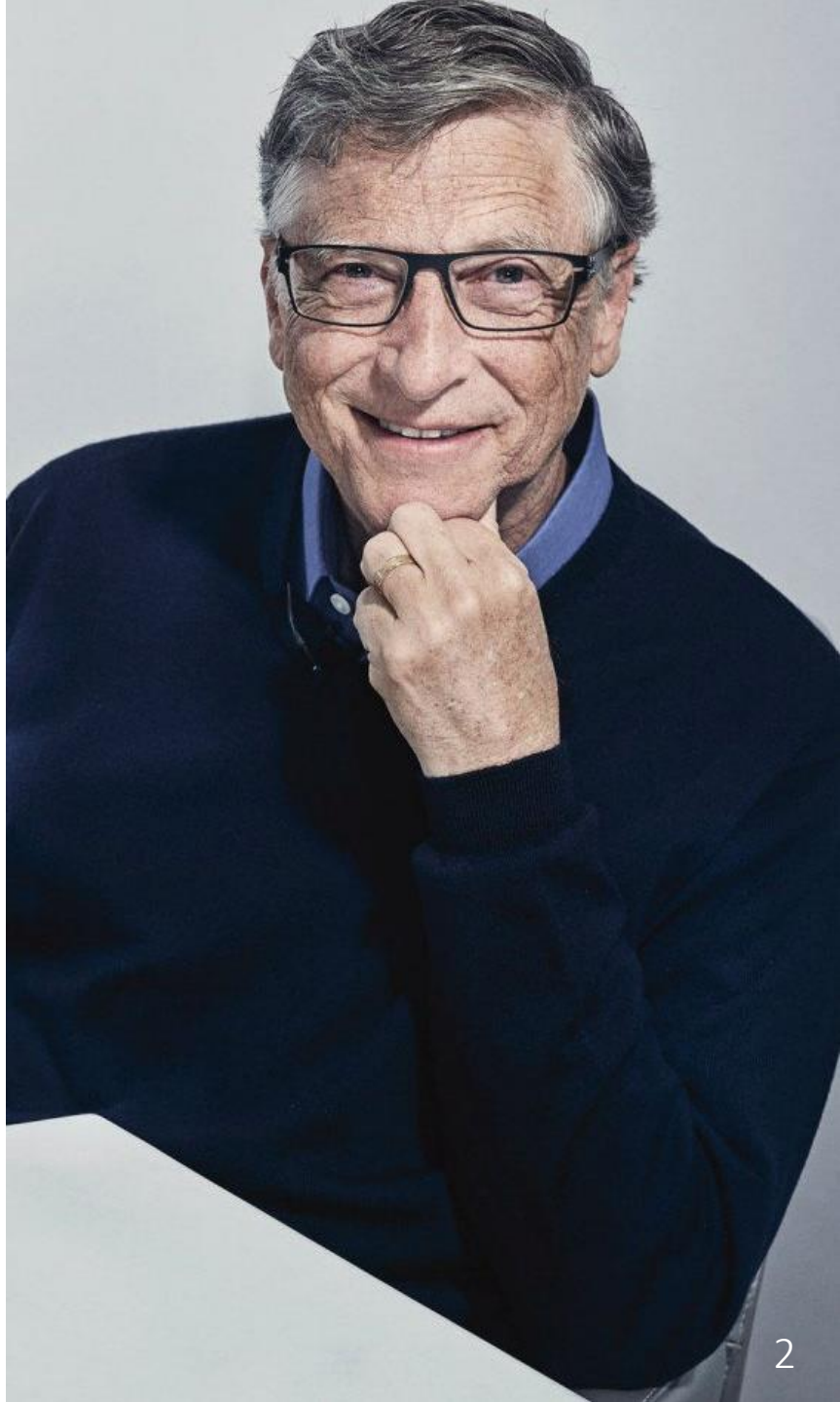
Научный руководитель: доц. Каф. «ВСТ», к.т.н. Гай В.Е.

Нижний Новгород

2020

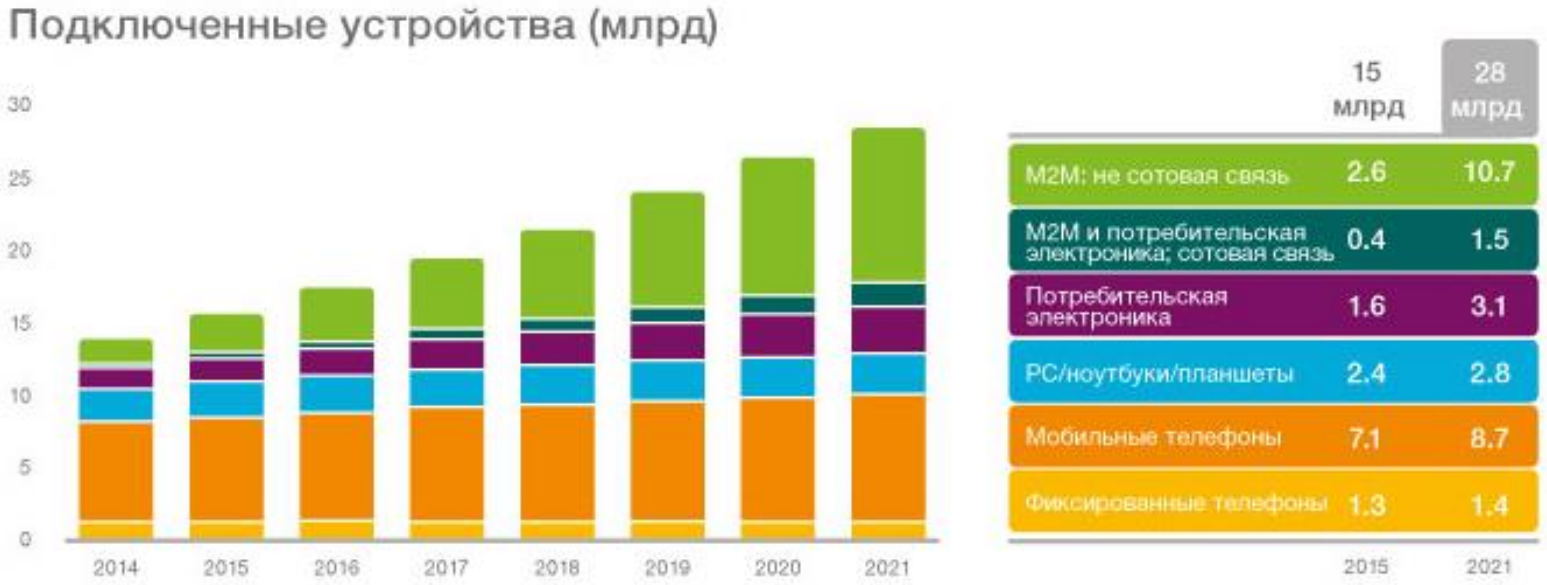
«В будущем на рынке останется два вида компаний: в Интернете и те, кто вышел из бизнеса»

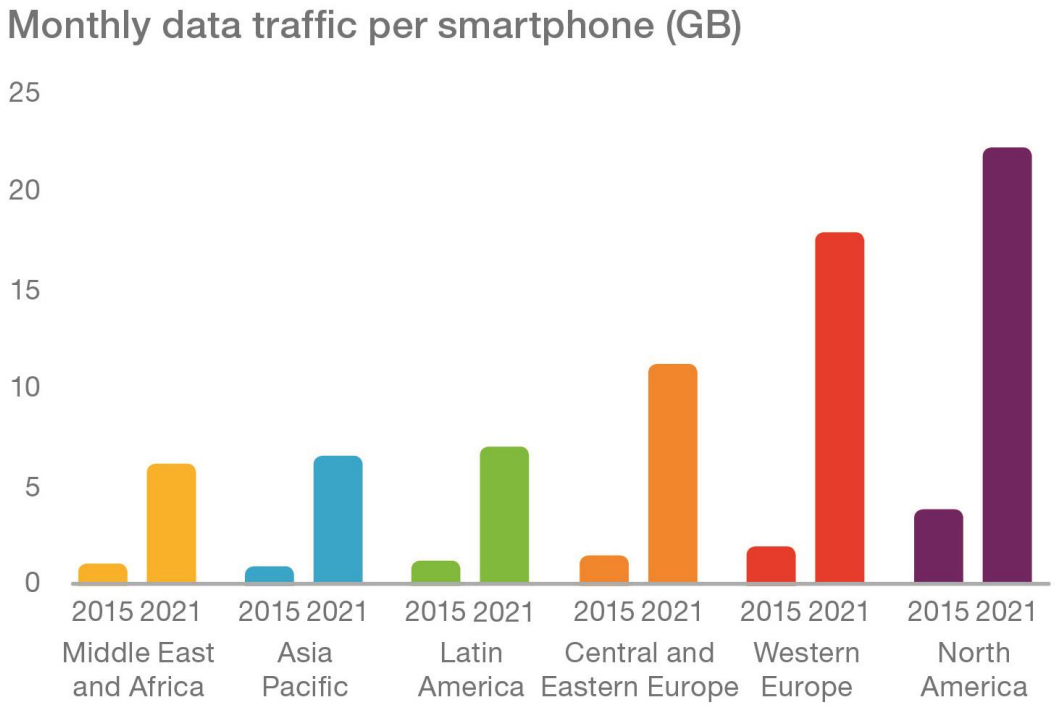
Билл Гейтс



Определение

Сетевой трафик – это объём информации, передаваемой через компьютерную сеть за определённый период времени.


Подключенные устройства (млрд)
30
25
20
15
10
5
0
2014
2015
2016
2017
2018
2019
2020
2021
15 млрд
28 млрд
M2M: не сотовая связь 2.6 10.7
M2M и потребительская электроника; сотовая связь 0.4 1.5
Потребительская электроника 1.6 3.1
PC/ноутбуки/планшеты 2.4 2.8
Мобильные телефоны 7.1 8.7
Фиксированные телефоны 1.3 1.4
2015
2021


Monthly data traffic per smartphone (GB)
25
20
15
10
5
0
2015 2021
Middle East and Africa
2015 2021
Asia Pacific
2015 2021
Latin America
2015 2021
Central and Eastern Europe
2015 2021
Western Europe
2015 2021
North America

Статистика

Научная новизна

Новый способ идентификации личности пользователя компьютера, заключающийся в статистическом анализе сетевого трафика на базе выделения поведенческих привычек пользователя и описания сетевых сессий с использованием алгоритмов классического машинного обучения.

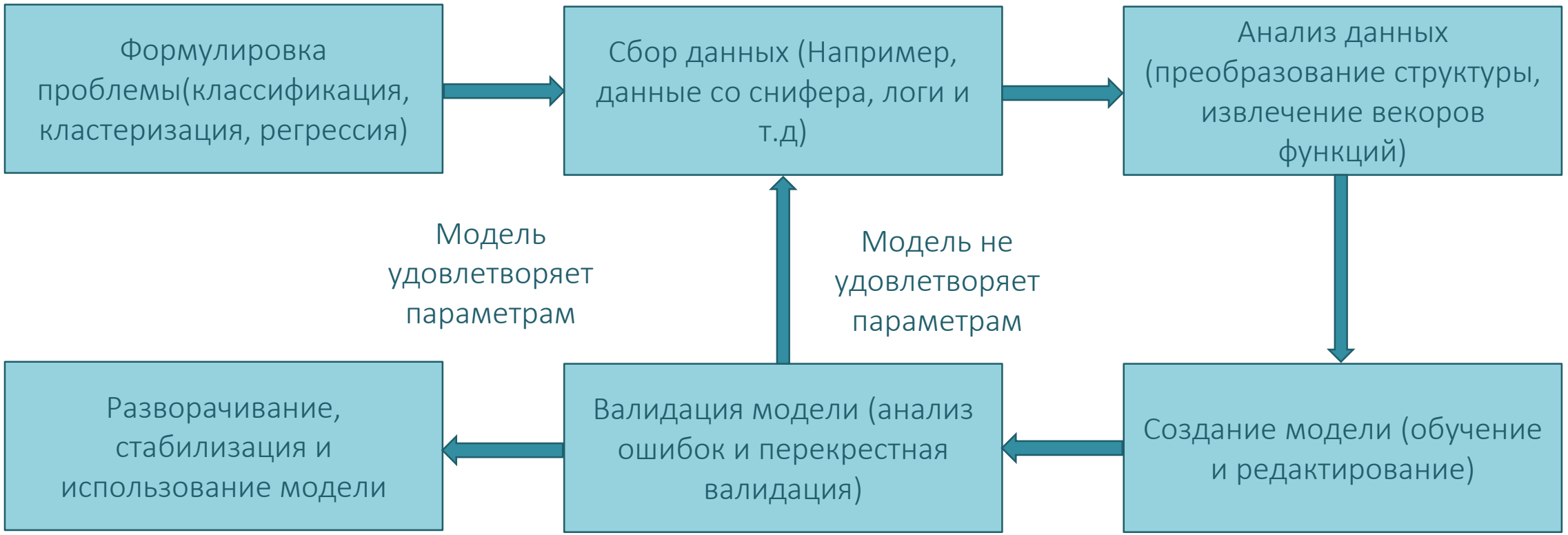
Цель исследования

Исследование различных методов классификации сетевого трафика, разработка модели идентификации пользователя по сетевому трафику с использованием алгоритмов машинного обучения и сравнение результатов.

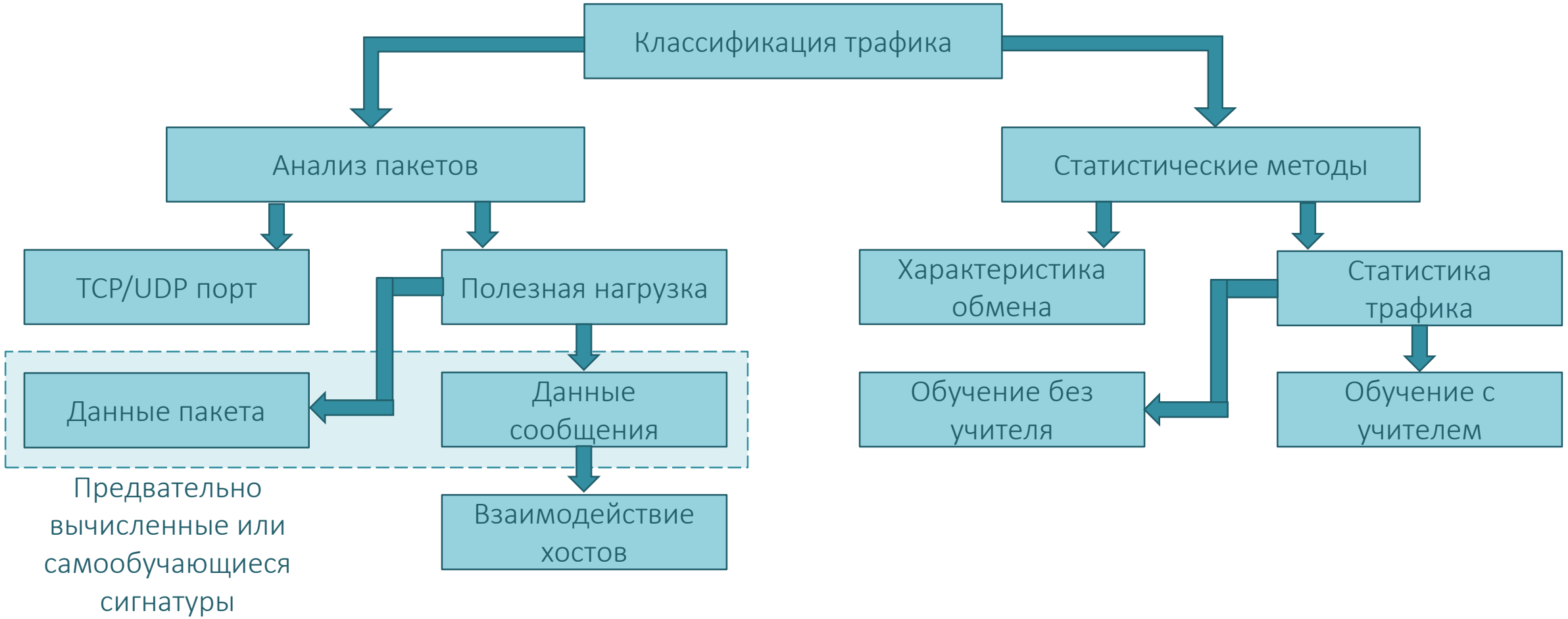
Задачи исследования

- [x] Обзор и анализ методов решения задачи
- [x] Сбор данных
- [x] Формирование признакового пространства
- [x] Создание приложения
- [x] Вычислительный эксперимент

Модель


Формулировка проблемы(классификация, кластеризация, регрессия)
Сбор данных (Например, данные со снифера, логи и т.д)
Анализ данных (преобразование структуры, извлечение векоров функций)
Модель удовлетворяет параметрам
Модель не удовлетворяет параметрам
Разворачивание, стабилизация и использование модели
Валидация модели (анализ ошибок и перекрестная валидация)
Создание модели (обучение и редактирование)

Методы решения задачи классификации сетевого трафика



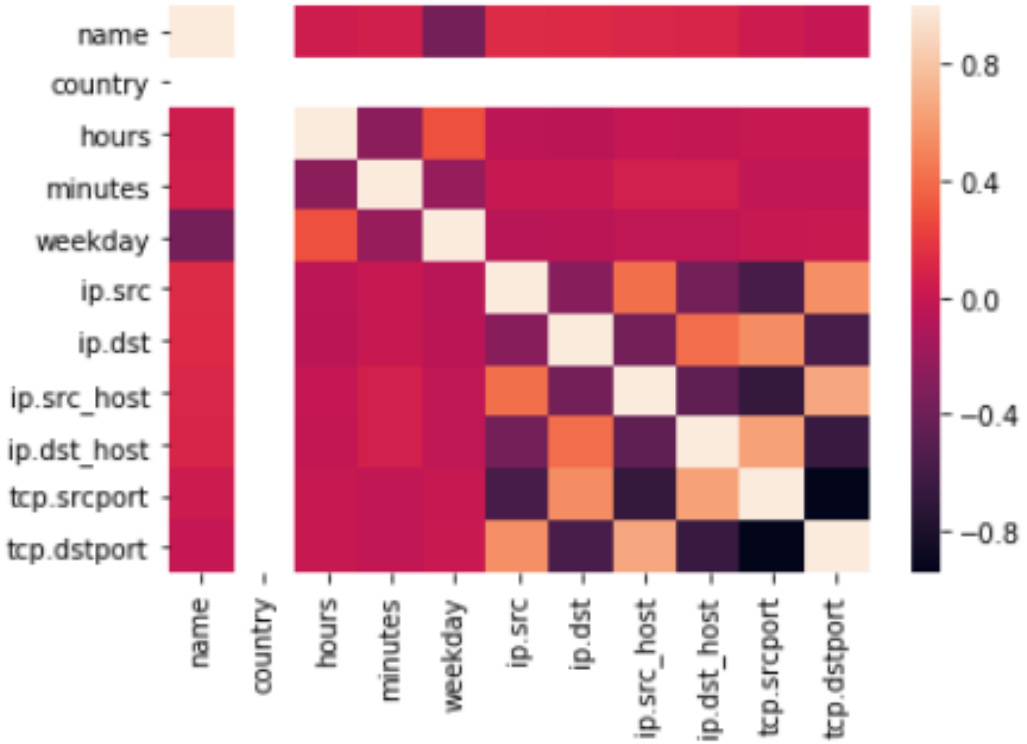
Сбор данных

- Захват сетевых пакетов на устройстве пользователя с использованием сниффера
- Получение статистики использования с маршрутизатора сети
- Межсетевой экран ПК-маршрутизатора
- Интерфейсы операционной системы или физические сетевые интерфейсы

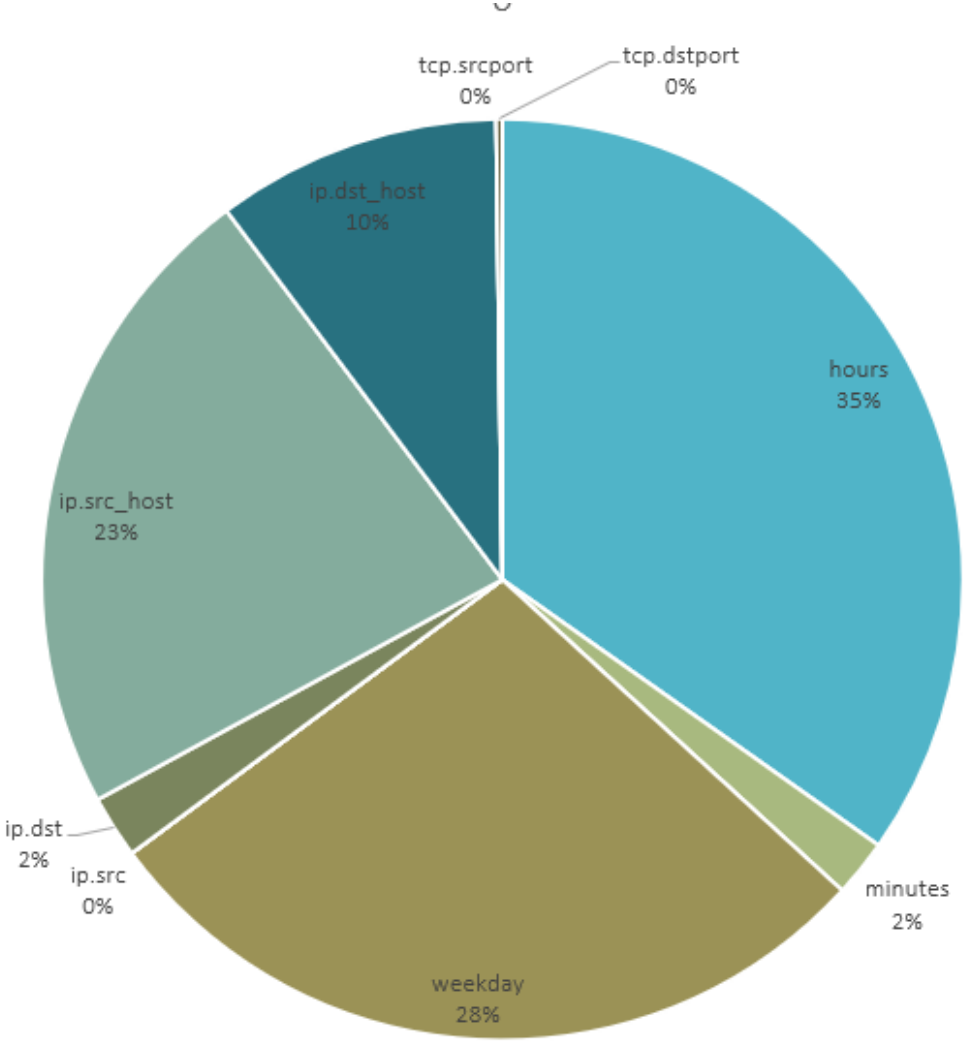
Анализ данных и формирование признакового пространства


```
{
    "_index": "packets-2020-02-07",
    "_type": "doc",
    "_score": null,
    "_source": {
    “layers”:  {
      “frame”: {…},
      “eth”: {…},
      “ip”: {…},
      “tcp”: {…},
      …
     }
}
```

```
    {
        "hours": 15,
        "minutes": 49,
        "country": “Russia",
        “user": “Olga",
        "weekday": 1,
        "ip.src": "13.83.151.160",
        "ip.dst": "192.168.1.195",
        "ip.src_host": "fe2cr.update.micr
osoft.com.akadns.net",
        "ip.dst_host": "192.168.1.195",
        "tcp.srcport": "443",
        "tcp.dstport": "50002"
    }
```

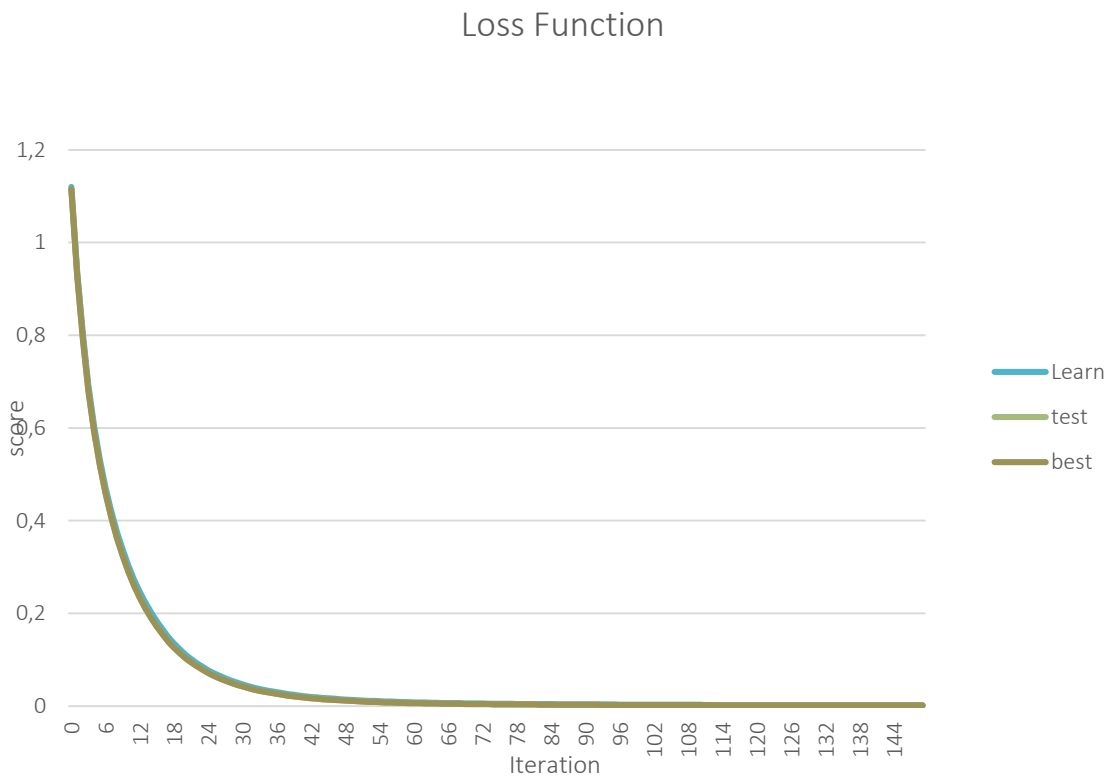


Результаты: матрица корреляции



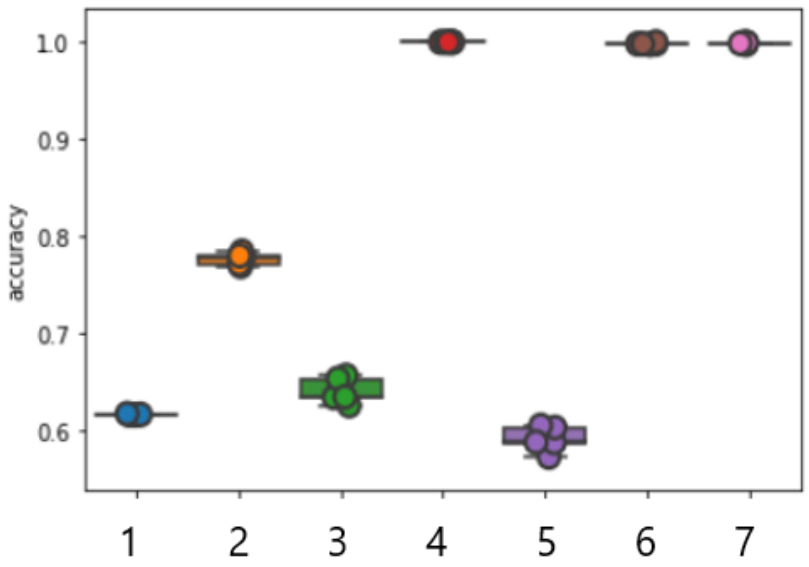
Результаты: влияние отдельных признаков на результат

Результаты



Class	Accuracy	Precision	Recall	F1	AUC	time
0	1.0	0,9897	0,9914	0,9944	0,9997	2m 37s
1	1.0	0,9979	0,9808	0,9863		
2	1.0	1.0	1.0	1.0		
3	1.0	0,9827	0,9827	0,9871		
AVG	1.0	0,9887	0,9887	0,9923		

Результаты



1. RandomForestClassifier
2. SVC
3. MiltinominalNB
4. DecisionTreeClassifier
5. KNNClassifier
6. LinearSVC
7. LogisticRegression

Classifier	Accuracy	Precision	Recall	F1	AUC	time
SVC	0.8434	0.87	0.84	0.83	0,84	0:01:28.053325
LinearSVC	0.9999	1.0	1.0	1.0	1.0	0:00:02.376799
LogRegression	1.0	1.0	1.0	1.0	1.0	0:00:00.117278
MultinomialNB	0.7954	0.9482	0.7954	0.855	0,9348	0:00:00.003857
KNN	0,7835	0.8652	0,7835	0,8035	0,8556	0:00:00.001470
DecisionTree	1.0	1.0	1.0	1.0	1.0	0:00:00.098747
RandomForest	0.6710	0,7689	0,6710	0,5813	0,9557	0:00:01.870440

Сравнение результатов с аналогами

Protocol	Precision	Recall
SSL	0.9513	0.9763
HTTP_Proxy	0.9174	0.9090
MySQL	0.9989	0.9993
SMB	1.0000	1.0000
HTTP_Connect	0.9967	0.9930
Whois-DAS	0.9943	0.9777
Redis	0.9985	0.9974
SSH	0.9996	1.0000
Apple	0.9640	0.9728
Kerberos	0.9996	0.9996
DCE_RPC	1.0000	1.0000
NetBIOS	1.0000	1.0000
FTP_CONTROL	0.9970	0.9973
DNS	0.9989	0.9985
Skype	0.9779	0.9722
LDAP	0.9996	0.9992
AppleiCloud	0.9679	0.9689
AppleiTunes	0.9520	0.9617
MSN	0.9453	0.9230
GMail	0.9953	0.9973
BitTorrent	0.9992	0.9992
TDS	1.0000	1.0000
IMAPS	0.9814	0.9654
SMTP	0.9949	0.9883
RSYNC	0.9987	0.9993

Zhanyi Wang: "The application of deep learning on traffic identification"

Tool	*Complexity (Signature)*	*Training size*	*Success rate*
DirBuster	low	10-50	high (99%)
Burp Suite	none (plain TCP)	5000+	low (40%)
Nessus	complex	5000+	medium (85%)
sqlmap	low	10-50	high (99%)
Nikto	low/medium	10-50	high (95%)

P.Fruhwirt, S. Schrittwieser, E.R. Weippl: "Using machine learning techniquesfor traffic classification and preliminary surveying of an attacker's profile"

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
J48	0,999	0,007	0,998	0,999	0,999	0,994	0,999	1,000	nonTOR
	0,993	0,001	0,997	0,993	0,995	0,994	0,999	0,997	TOR
Weighted Avg.	0,998	0,006	0,998	0,998	0,998	0,994	0,999	0,999	
J48Consolidated	0,998	0,002	0,999	0,998	0,999	0,993	1,000	1,000	nonTOR
	0,998	0,002	0,991	0,998	0,994	0,993	1,000	0,998	TOR
Weighted Avg.	0,998	0,002	0,998	0,998	0,998	0,993	1,000	1,000	
BayesNet	0,982	0,020	0,996	0,982	0,989	0,938	0,999	1,000	nonTOR
	0,980	0,018	0,918	0,980	0,948	0,938	0,999	0,995	TOR
Weighted Avg.	0,982	0,020	0,982	0,982	0,982	0,938	0,999	0,999	
jRip	1,000	0,000	1,000	1,000	1,000	0,999	1,000	1,000	nonTOR
	1,000	0,000	0,998	1,000	0,999	0,999	1,000	0,999	TOR
Weighted Avg.	1,000	0,000	1,000	1,000	1,000	0,999	1,000	1,000	
OneR	0,997	0,000	1,000	0,997	0,999	0,992	0,999	1,000	nonTOR
	1,000	0,003	0,987	1,000	0,994	0,992	0,999	0,987	TOR
Weighted Avg.	0,998	0,000	0,998	0,998	0,998	0,992	0,999	0,997	
RepTREE	0,987	0,032	0,997	0,987	0,992	0,940	0,999	0,998	nonTOR
	0,982	0,019	0,923	0,983	0,951	0,941	0,998	0,997	TOR
Weighted Avg.	0,983	0,021	0,984	0,983	0,984	0,939	0,998	0,998	

Alfredo Cuzzocrea, Fabio Martinelli, Francesco Mercaldo, Gianni Vercelli: "TorTraffic Analysis and Detection via Machine Learning Techniques"

Заключение

Q&A