Федеральное государственное бюджетное образовательное
учреждение высшего образования
«МИРЭА – Российский Технологический Университет»

Разработка алгоритма обнаружения спама с помощью методов машинного обучения

Выполнила студентка
группы ББМО-01-21
Концевая Юлия Марковна

Москва – 2021

Алгоритмы

- работает плохо, в случае, когда признаков очень много

Метод k ближайших соседей

- "проблемой нулевой частоты"

Наивный байесовский классификатор

- склонность к переобучению

Деревья решений

- требует больше времени для обучения

Метод опорных векторов

- требует больше вычислительной мощности

Случайный лес

- требует большого набора данных

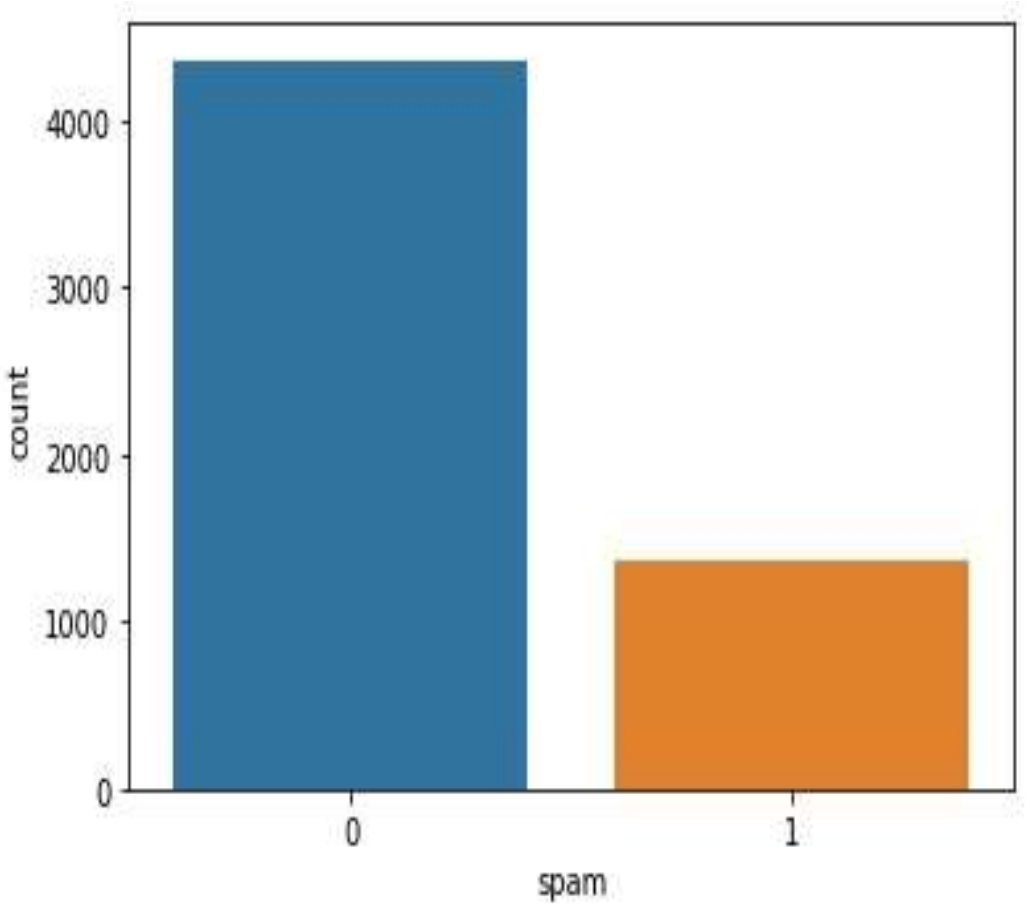
Логистическая регрессия

Данные для обучения

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1
...	...	...
5723	Subject: re : research and development charges...	0
5724	Subject: re : receipts from visit jim , than...	0
5725	Subject: re : enron case study update wow ! a...	0
5726	Subject: re : interest david , please , call...	0
5727	Subject: news : aurora 5 . 2 update aurora ve...	0

5728 rows × 2 columns

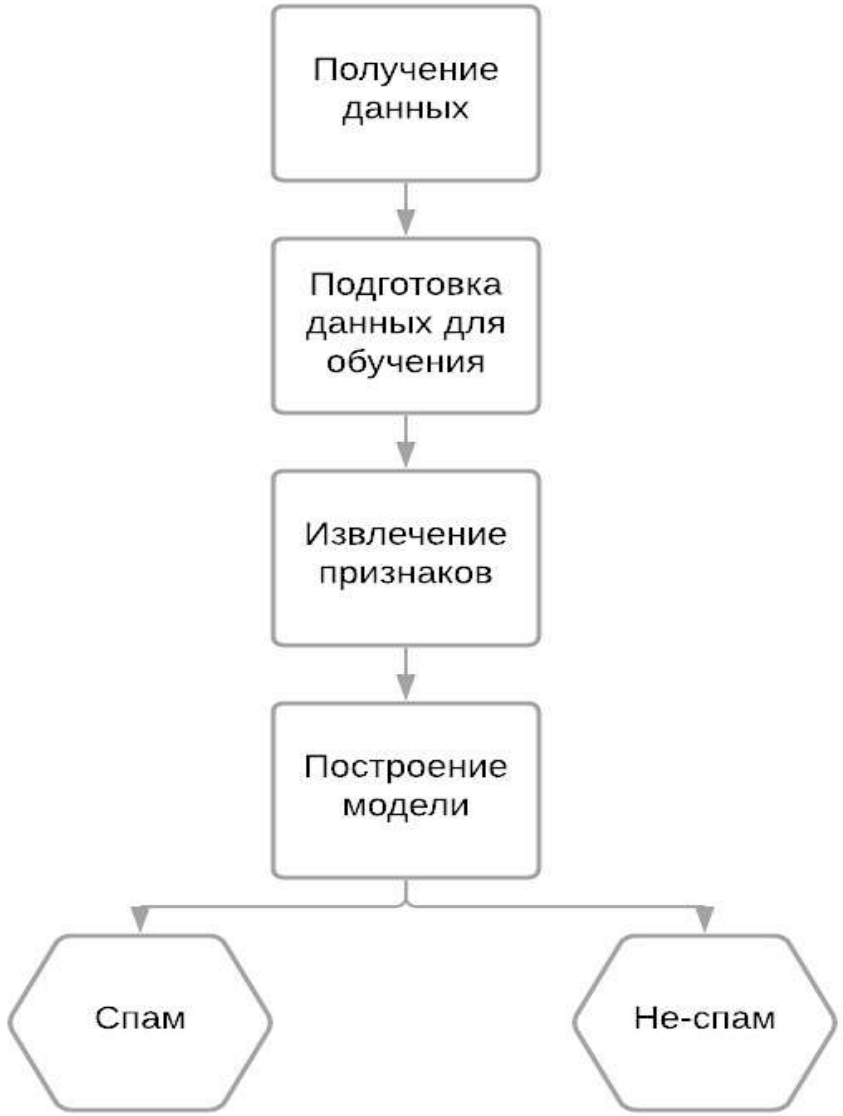
Обучающая выборка



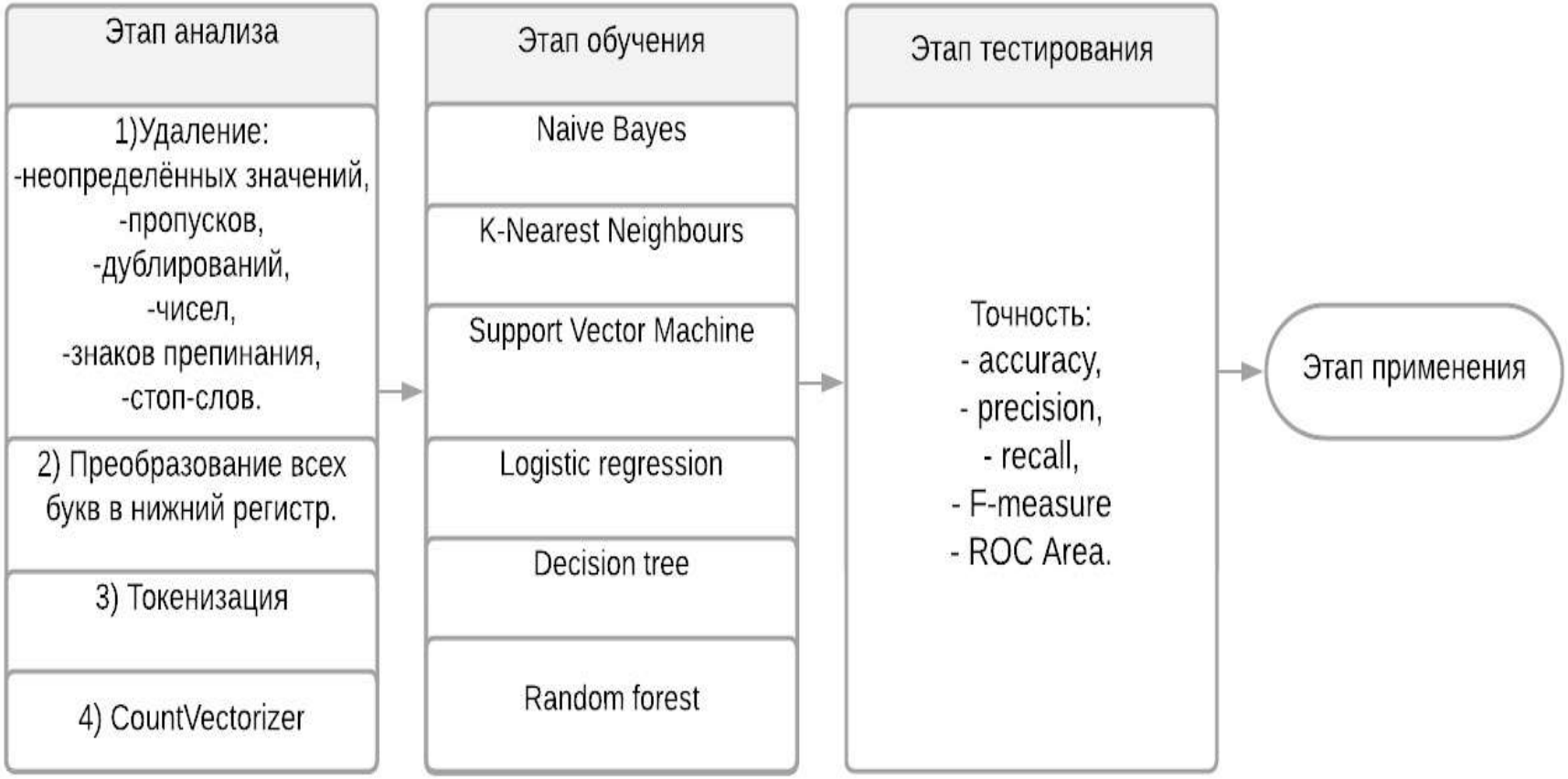
Распределение спам и не-спам сообщений

Общая схема решаемой задачи

- Язык программирования – **Python**;
- Среда - Jupyter Notebook.



Этапы разработки алгоритма



Этап тестирования

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

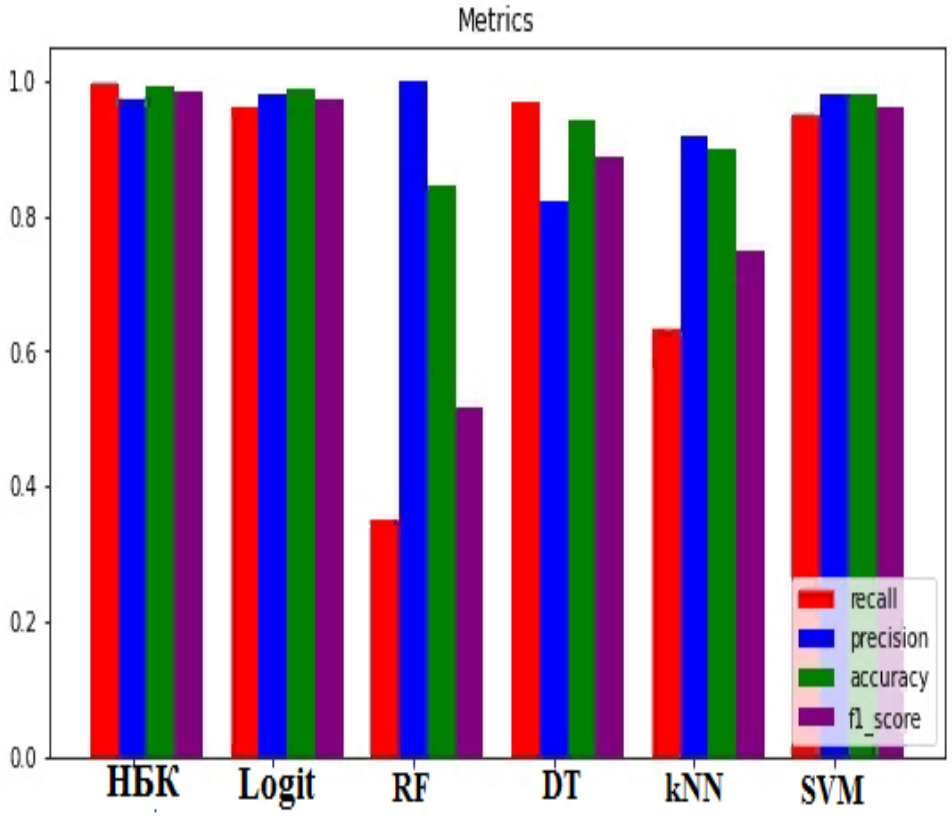
$$\text{accuracy} = \frac{TN+TP}{TP+FN+FP+TP}$$

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

$$\text{F-value} = \frac{2*precision*recall}{pecision+recall}$$

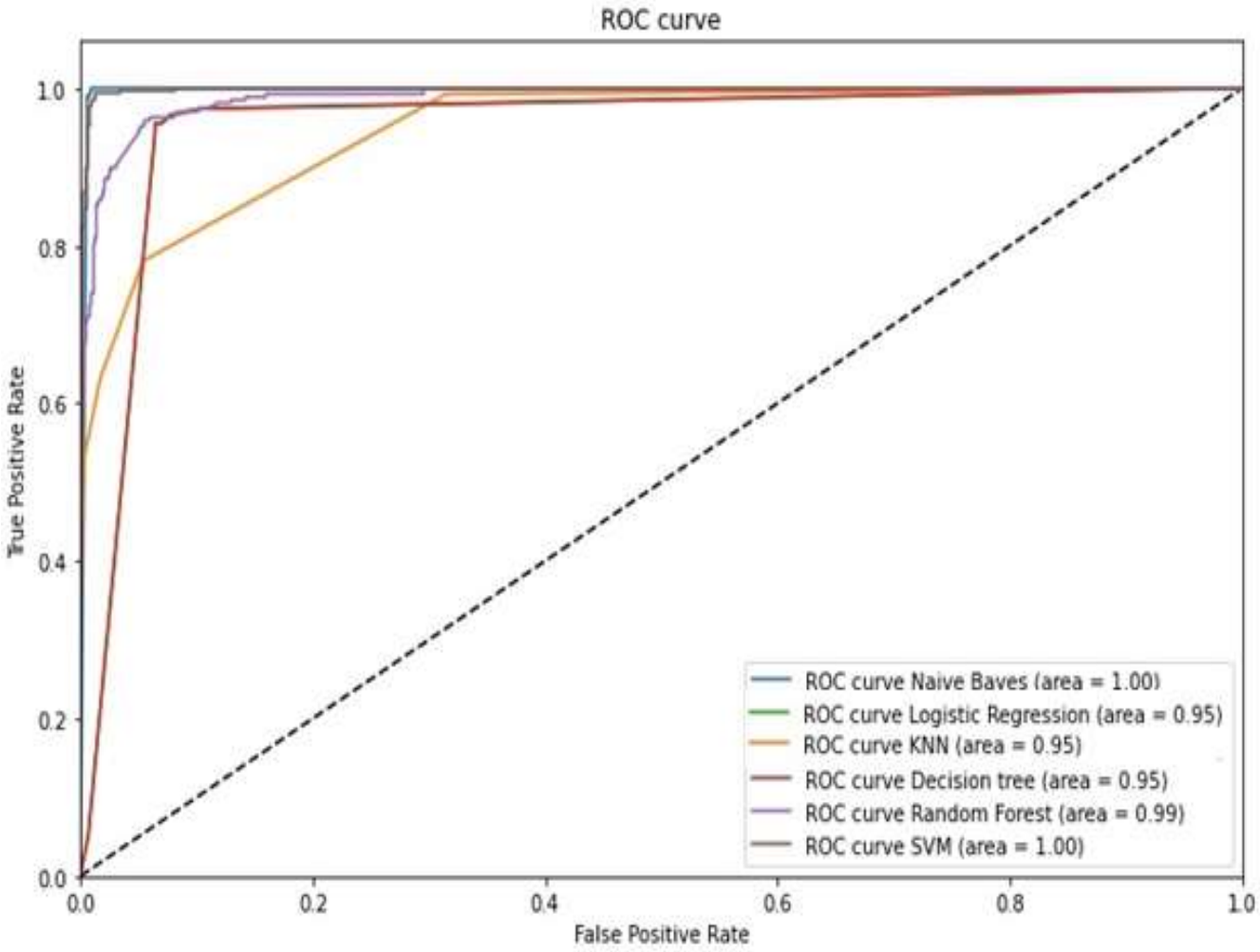
Оценка качества моделей



Точность алгоритмов

Алгоритм	Accuracy	Precision	Recall	F-measure	ROC Area
kNN	**0,90**	0,91	0,63	0,74	0,95
НБК	**0,99**	0,97	0,99	0,98	1.00
Деревья решений	**0,94**	0,82	0,96	0,88	0,95
SVM	**0,98**	0,98	0,95	0,96	1.00
Случайный лес	**0,84**	1	0,28	0,42	0,99
Логистическая регрессия	**0,99**	0,98	0,96	0,97	0,95

Анализ классификации на основе ROC-кривых



Логистическая регрессия даёт наилучшие результаты классификации.

Федеральное государственное бюджетное образовательное
учреждение высшего образования
«МИРЭА – Российский Технологический Университет»

Разработка алгоритма обнаружения спама с помощью методов машинного обучения

Выполнила студентка
группы ББМО-01-21
Концевая Юлия Марковна

Москва – 2021