

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ
ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ»

Кафедра «Измерительно-вычислительные комплексы»

«Методы искусственного интеллекта»

Отчёт по лабораторной работе №5

Выполнил:

студент группы ИСТбд-42

Миридонов Павел

Проверил:

доцент кафедры ИВК, к.т.н.

Шишкин В.В

Ульяновск
2023

Задание на лабораторную работу:

1. Ознакомиться с классификаторами библиотеки Scikit-learn
 2. Выбрать для исследования не менее 3 классификаторов
 3. Выбрать набор данных для задач классификации из открытых источников
<https://tproger.ru/translations/the-best-datasets-for-machine-learning-and-data-science/>
<https://vc.ru/ml/150241-15-proektov-dlya-razvitiya-navykov-raboty-s-mashinnym-obucheniem>
<https://archive.ics.uci.edu/ml/index.php>
<https://habr.com/ru/company/edison/blog/480408/>
<https://www.kaggle.com/datasets/>
- учебные наборы библиотеки Scikit-learn
4. Выбор классификаторов и набора данных утвердить у преподавателя (не должно быть полного совпадения с выбором другого студента)
 5. Для каждого классификатора определить целевой столбец и набор признаков. Обосновать свой выбор. При необходимости преобразовать типы признаков данных.
 6. Подготовить данные к обучению.
 7. Провести обучение и оценку моделей на сырых данных.
 8. Провести предобработку данных.
 9. Провести обучение и оценку моделей на очищенных данных.
 10. Проанализировать результаты.
 11. Результаты анализа представить в табличной и графической форме.
 12. Сформулировать выводы.
 13. Оформить отчет по л/р.
 14. Защитить результаты работы.

1. Для исследования были выбраны следующие классификаторы:
 - 1) К ближайших соседей
 - 2) Случайный лес
 - 3) Наивный байесовский метод
2. Был выбран набор данных, в котором отражен химический состав воды и ее пригодность для питья
<https://www.kaggle.com/datasets/adityakadiwal/water-potability>
3. В качестве набора признаков были использованы:
 - 1) pH воды
 - 2) Жесткость воды (Hardness)
 - 3) Минерализация (solids)
 - 4) Хлорамины
 - 5) Сульфаты
 - 6) Проводимость воды
 - 7) Органический углерод
 - 8) Тригалометаны
 - 9) Мутность воды
4. За целевой столбец для каждого классификатора был принят столбец, который отражает пригодность воды для питья, потому что он является выходным для датасета.

Результат работы программы:

Классификатор КНН:

[[629 212]

[343 224]]

True positive: 224

True negative: 629

False positive: 212

False negative: 343

	precision	recall	f1-score	support
0	0.65	0.75	0.69	841
1	0.51	0.40	0.45	567
accuracy			0.61	1408
macro avg	0.58	0.57	0.57	1408
weighted avg	0.59	0.61	0.59	1408

Точность модели: 0.6058238636363636

Классификатор Random Forest:

[[710 125]

[368 205]]

True positive: 205

True negative: 710

False positive: 125

False negative: 368

	precision	recall	f1-score	support
0	0.66	0.85	0.74	835
1	0.62	0.36	0.45	573
accuracy			0.65	1408
macro avg	0.64	0.60	0.60	1408
weighted avg	0.64	0.65	0.62	1408

Точность модели: 0.6498579545454546

Классификатор Naive Bayes:

[[731 92]

[467 118]]

True positive: 118

True negative: 731

False positive: 92

False negative: 467

	precision	recall	f1-score	support
0	0.61	0.89	0.72	823
1	0.56	0.20	0.30	585
accuracy			0.60	1408
macro avg	0.59	0.54	0.51	1408
weighted avg	0.59	0.60	0.55	1408

Точность модели: 0.6029829545454546

Вывод графиков:

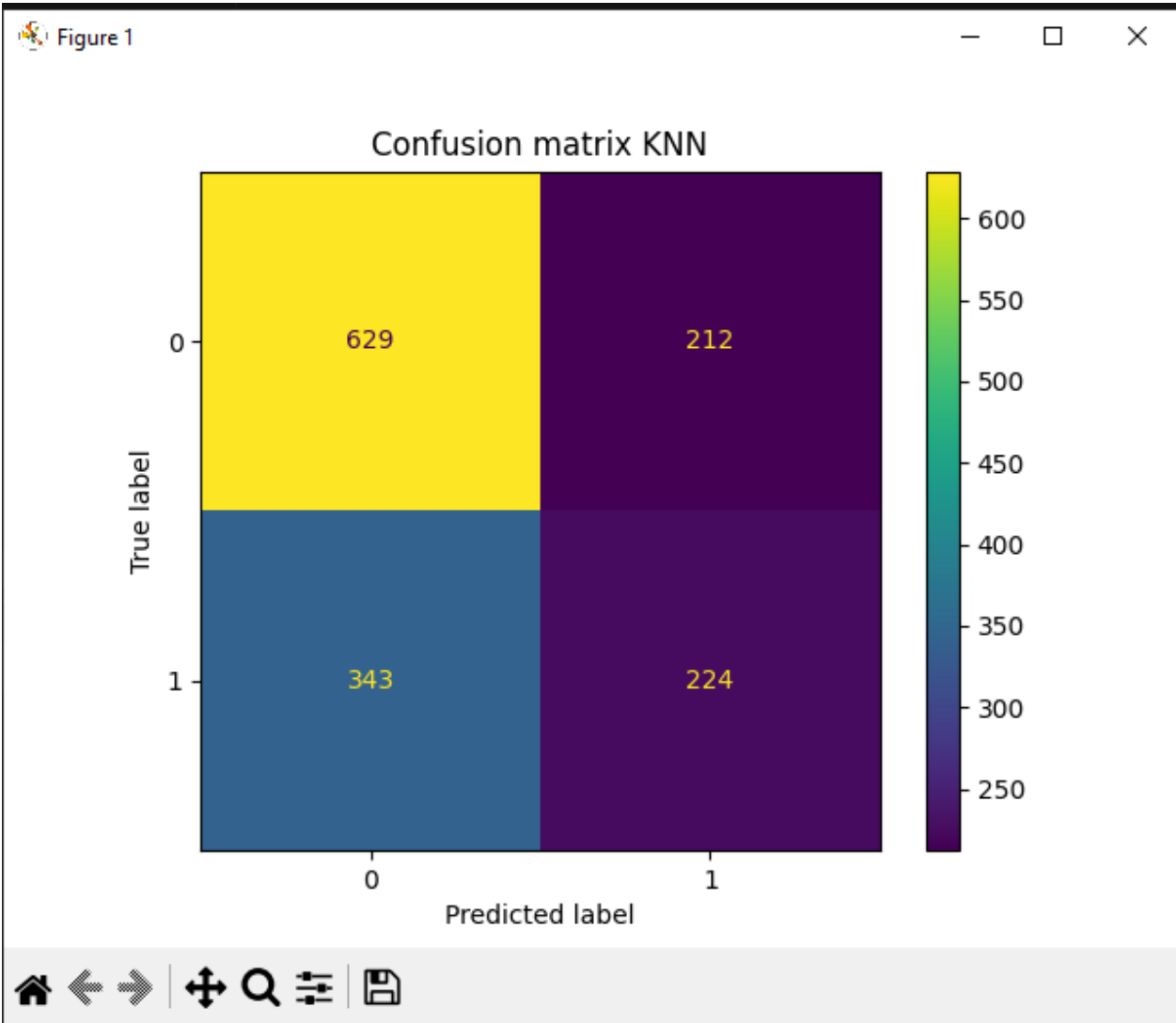
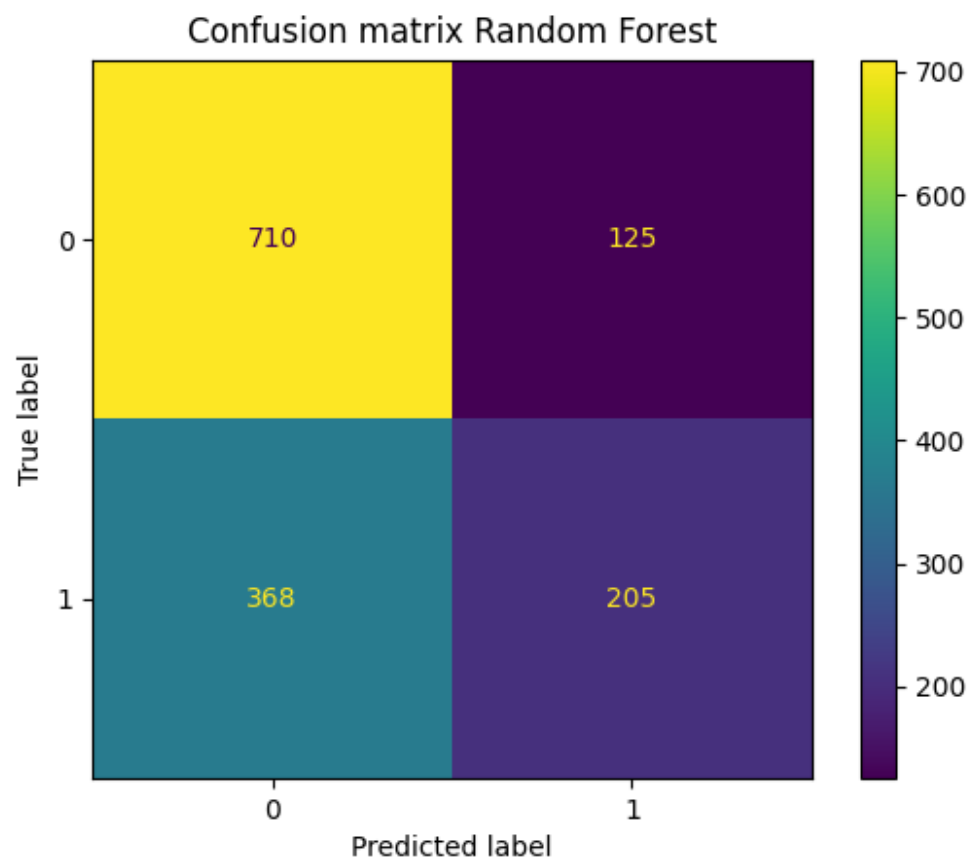
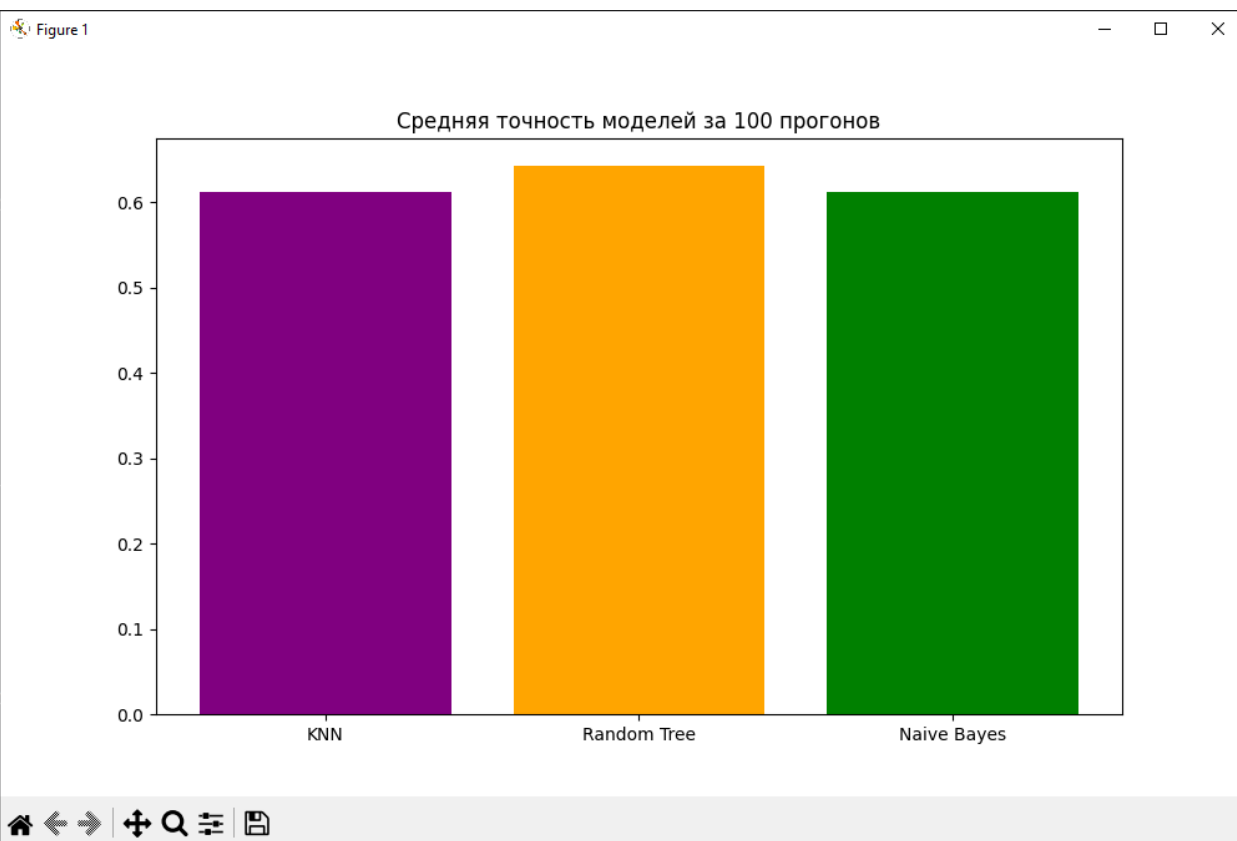
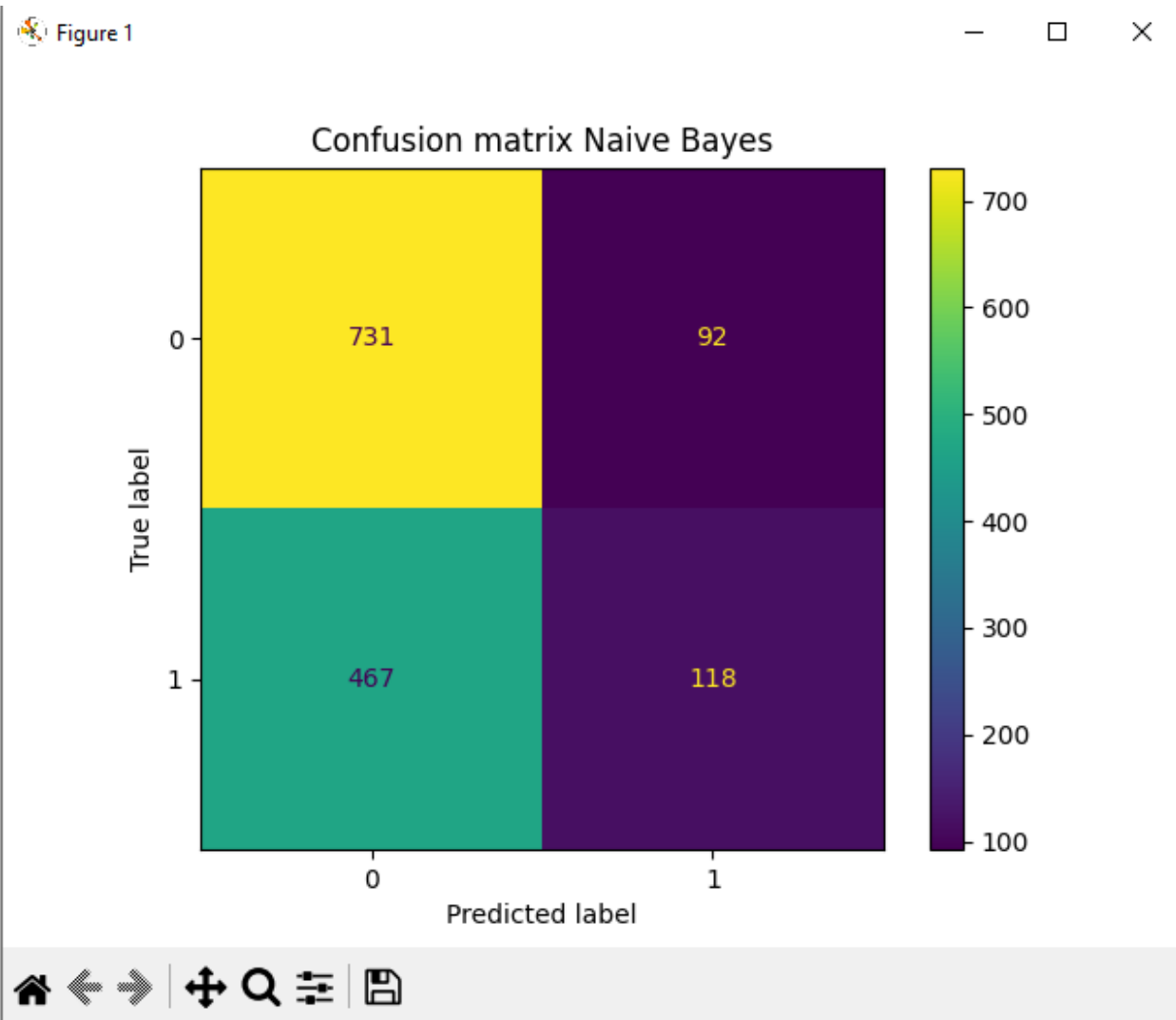


Figure 1





Вывод: в результате выполнения лабораторной работы были изучены следующие классификаторы библиотеки sklearn: K ближайших соседей, случайный лес, наивный байесовский метод. По итогам тестирования наиболее точной моделью оказалась модель, в которой использовался классификатор случайный лес. Но разница в точности оказалась не особо большой, возможно, это связано с достаточно большим размером выборки.