|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Технологии и инструментарий анализа больших данных»

**Практическая работа № 10**

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнил студент группы ИНБО-01-17 | ИВБО-06-20, Чурилов А.К. | (подпись) | |
| Принял | Парамонов А.А. | (подпись) | |
| Отчет представлен | «17» ноября 2023 г. | |  | |

Москва 2023 г.

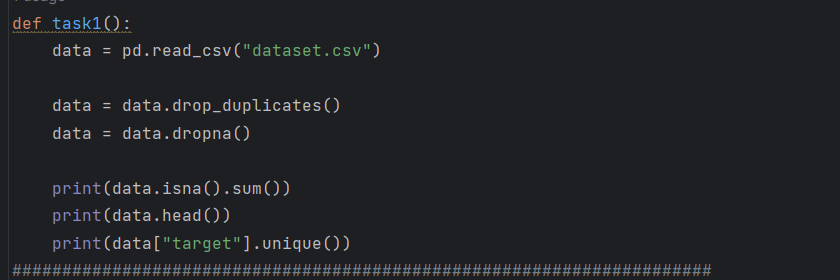
**Задание**

1. Найти данные для классификации. Данные в группе повторяться не должны! Предобработать данные, если это необходимо.
2. Изобразить гистограмму, которая показывает баланс классов. Сделать выводы.
3. Разбить выборку на тренировочную и тестовую. Тренировочная для обучения модели, тестовая для проверки ее качества.
4. Применить алгоритмы классификации: логистическая регрессия, SVM, KNN. Построить матрицу ошибок по результатам работы моделей (использовать confusion\_matrix из sklearn.metrics).
5. Сравнить accuracy, precision, recall и f1-меру (можно использовать classification\_report из sklearn.metrics). Также сравнить время работы алгоритмов. Сделать выводы.
6. Оформить отчет о проделанной работе.

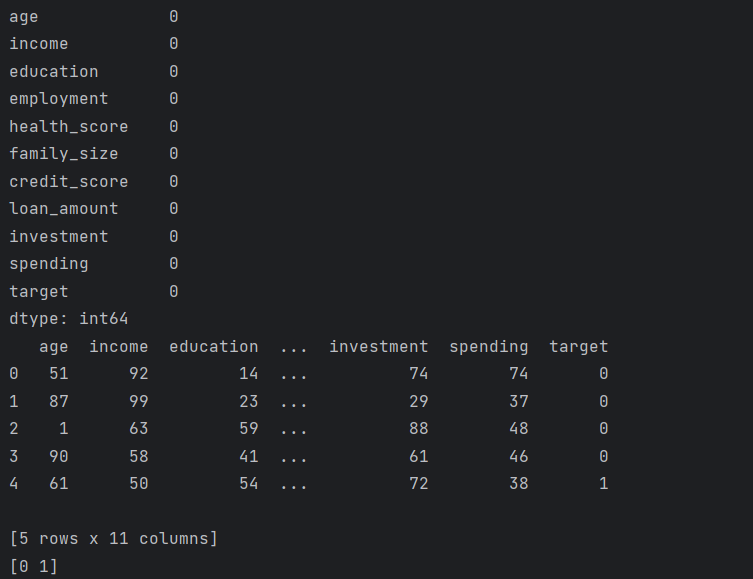
**Ход работы**

**Задание 1**

Найти данные для классификации. Данные в группе повторяться не должны! Предобработать данные, если это необходимо.

****

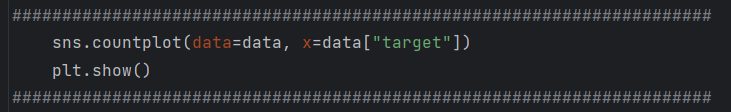
**Рисунок 1 – Задание 1**

****

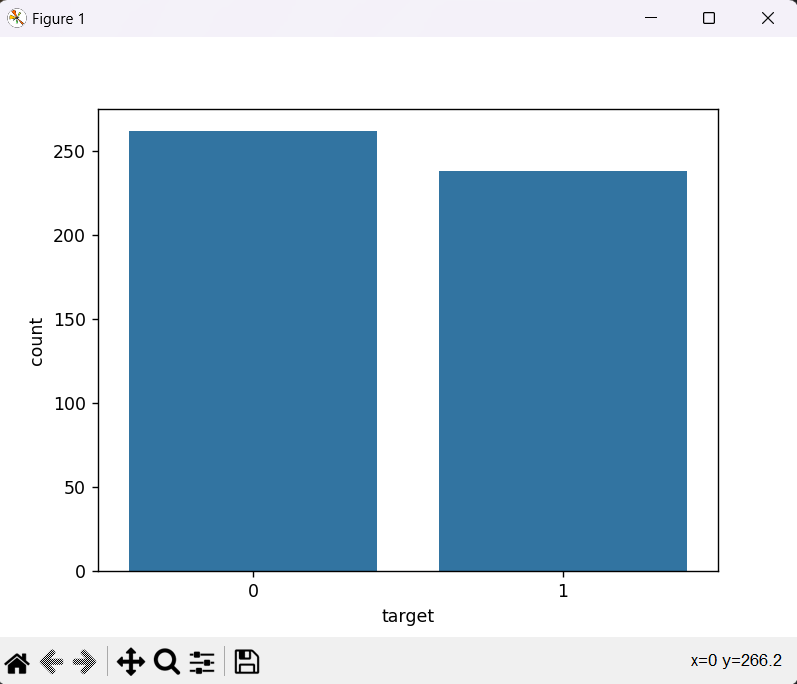
**Рисунок 2 – Результат задания 1**

Для выполнения практической работы использовался датасет с данными о сердечных приступах. Датасет был загружен с платформы Kaggle.

**Задание 2**

Изобразить гистограмму, которая показывает баланс классов. Сделать выводы. 

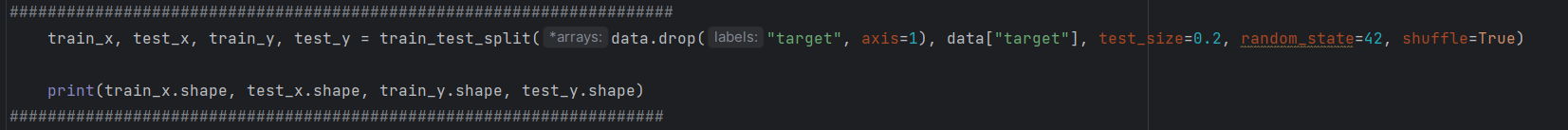
**Рисунок 3 – Задание 2**

****

По гистограмме можно сделать вывод, что классы достаточно сбалансированы, а значит проблем с этим быть не должно.

**Задание 3**

Разбить выборку на тренировочную и тестовую. Тренировочная для обучения модели, тестовая для проверки ее качества.

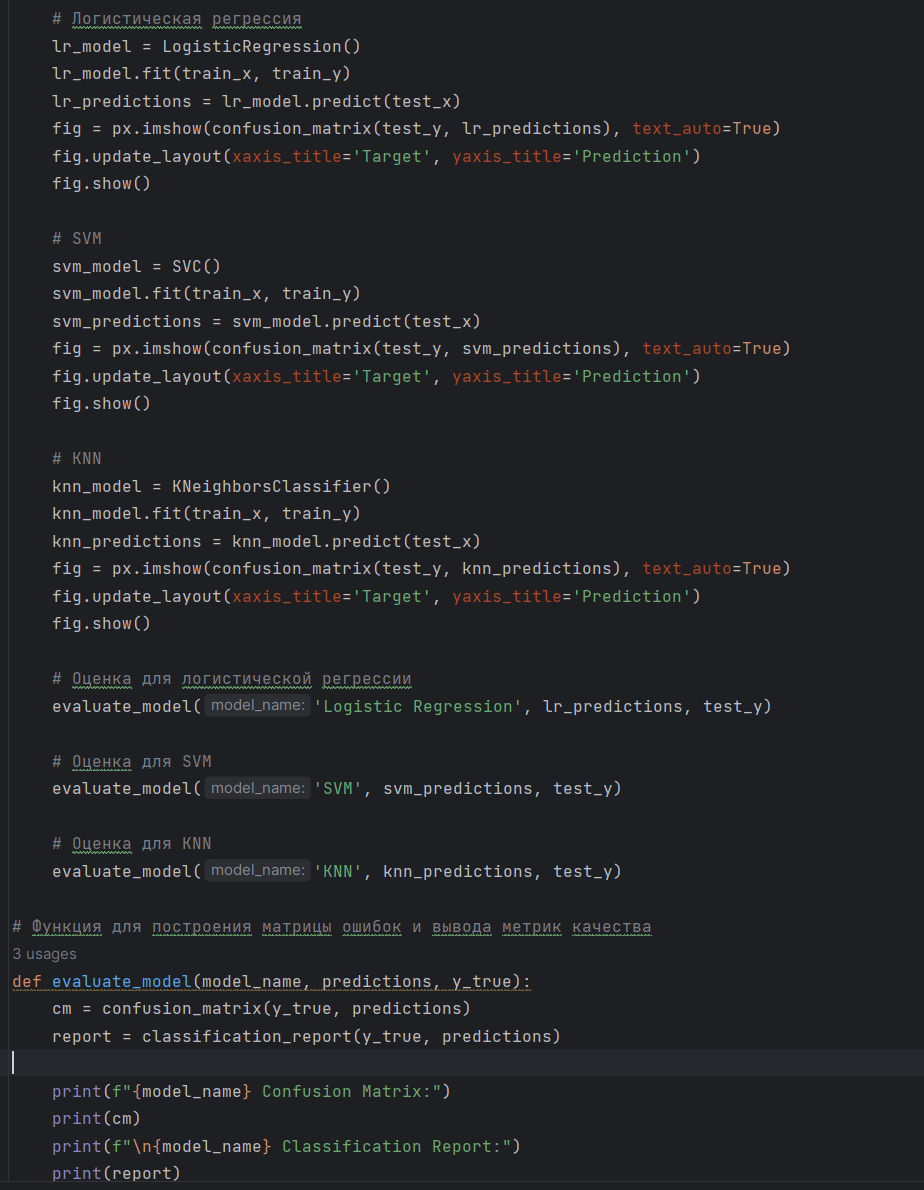
****

**Рисунок 4 – Задание 3**

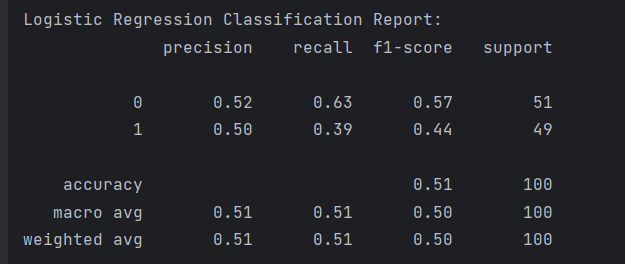
****

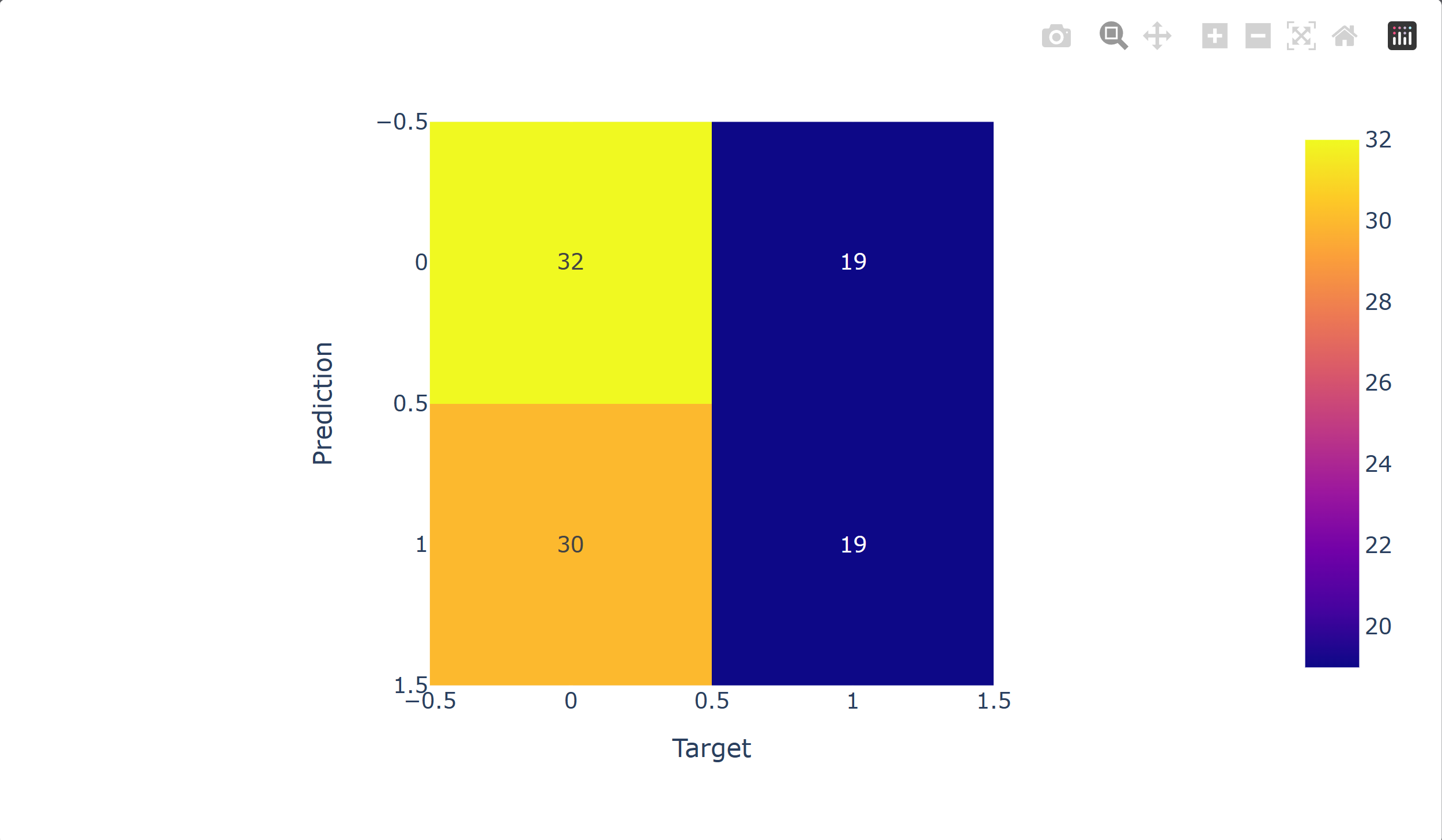
**Задание 4,5**

Применить алгоритмы классификации: логистическая регрессия, SVM, KNN. Построить матрицу ошибок по результатам работы моделей (использовать confusion\_matrix из sklearn.metrics). Сравнить accuracy, precision, recall и f1-меру (можно использовать classification\_report из sklearn.metrics). Сделать выводы.

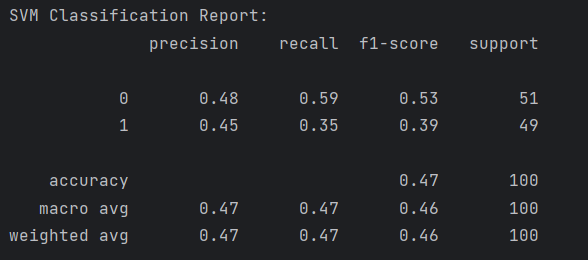
****

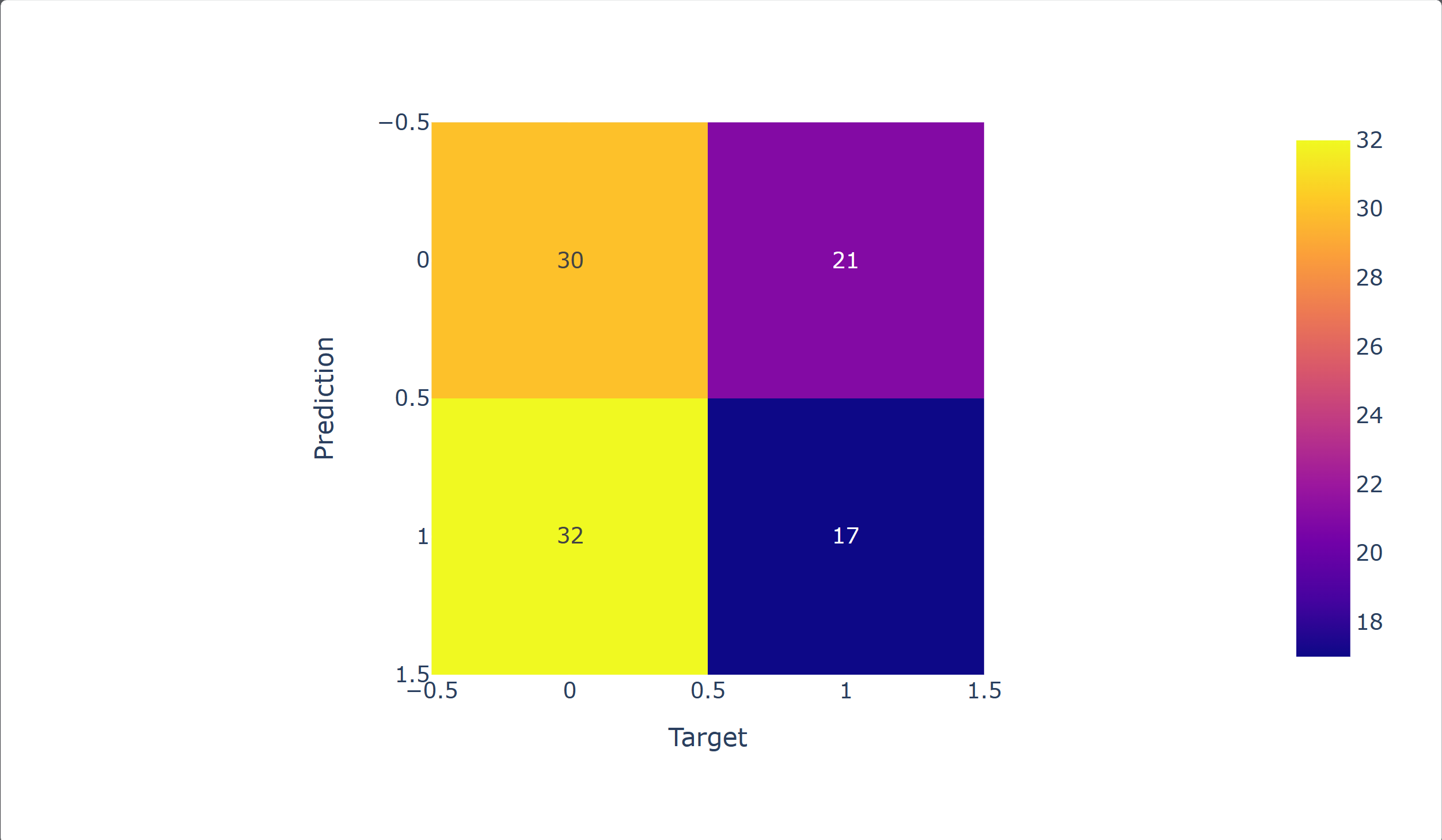
**Рисунок 5 – Задание 4**

****

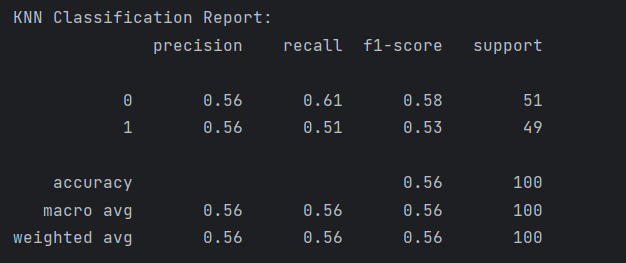
****

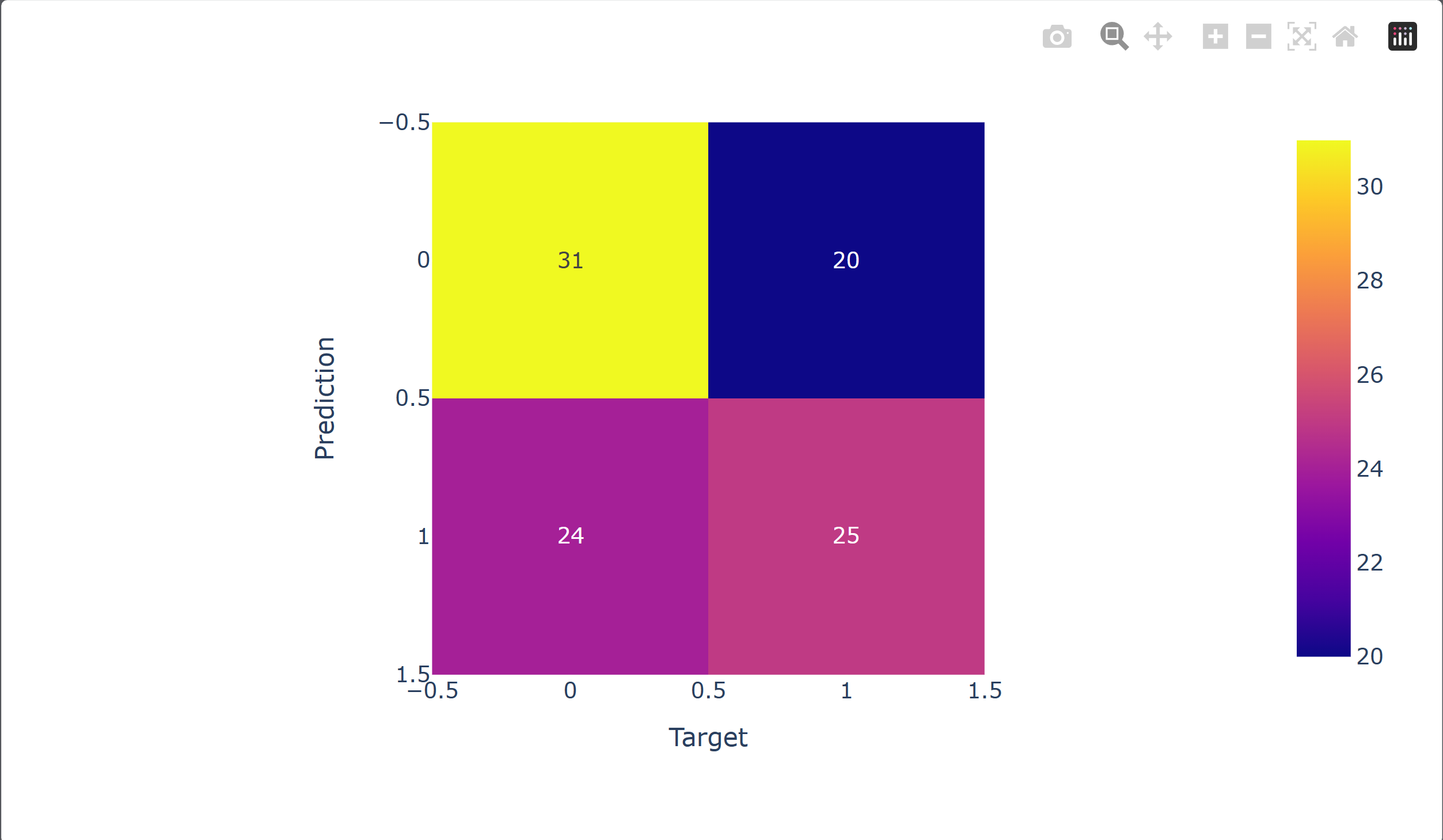
**Рисунок 6 – Результат задания 4,5, логистическая регрессия**

****

****

**Рисунок 7 – Результат задания 4,5, SVM**

****

****

**Рисунок 8 – Результат задания 4,5, KNN**

Можно сделать вывод, что обучается быстрее всего KNN, что связано с тем, что во время обучения данный алгоритм просто сохраняет обучающую выборку в отличии от остальных алгоритмов, которые производят расчёт оптимальных параметров. По точности – все алгоритмы отработали примерно одинаково, но точнее всего был алгоритм логистической регрессии.

**Вывод**

В ходе выполнения данной практической работы мы познакомились с алгоритмами классификации, способами их оценки, а также применили их на практике, сравнив качество и время обучения и предсказания.