МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

''УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ''

Кафедра «Информационно-вычислительные комплексы»

«Методы искусственного интеллекта»

Отчет по лабораторной работе №5

**Выполнила**:

студентка гр. ИСТбд-41

Кашицына Н. А.

**Проверил**:

преподаватель

Шишкин В. В

Ульяновск, 2022

**Исследование инструментов классификации библиотеки Scikit-learn**

**Задание:**

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 признаковых 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.

**Результат:**

Dataset - fast food chains, классификаторы - SGD Classifier, Support Vector Machines, Logistic Regression

Столбец Калории

0 660.0

1 740.0

2 790.0

3 900.0

4 980.0

...

445 280.0

446 330.0

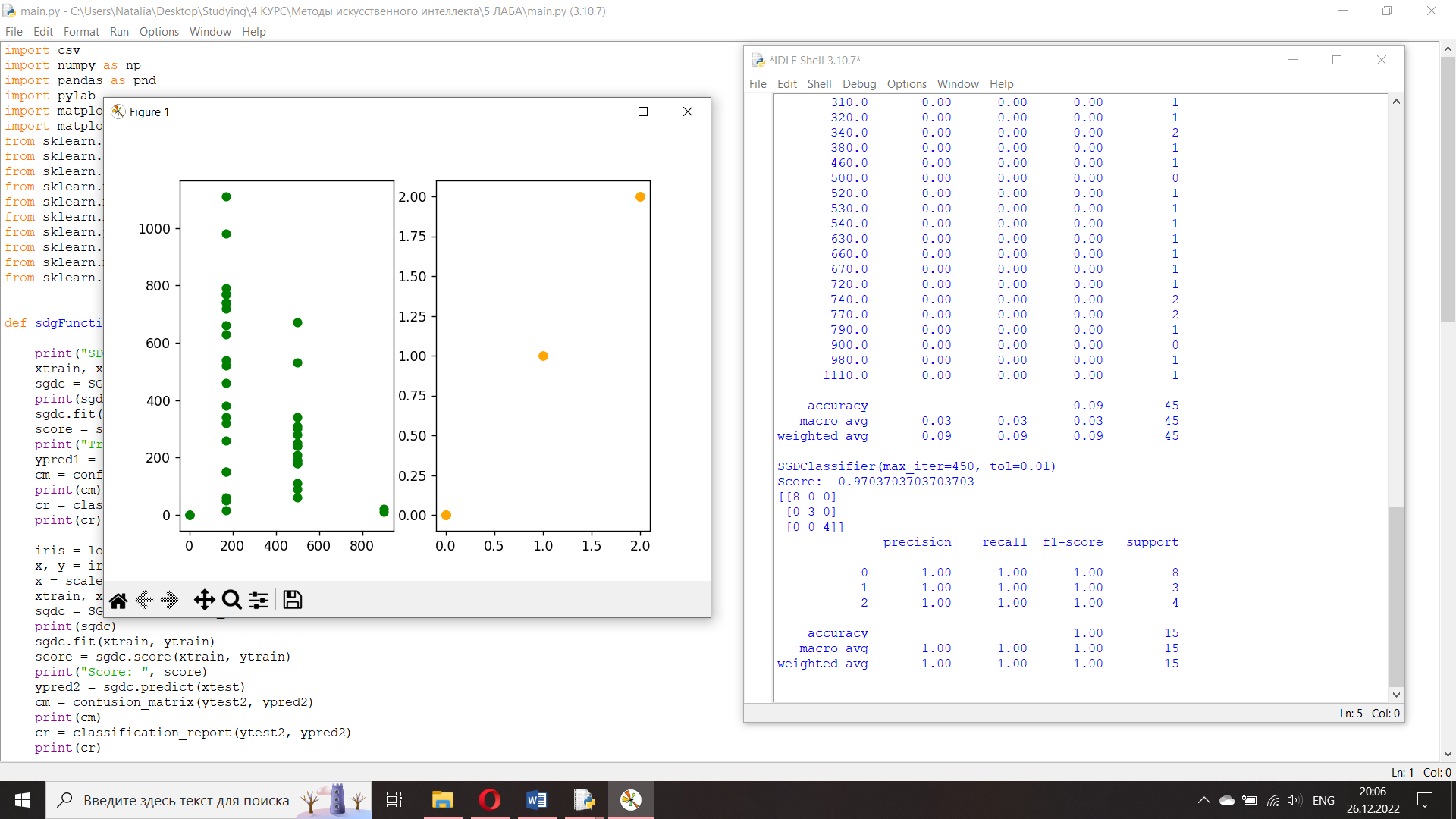
447 200.0

448 240.0

449 290.0

Name: Calories, Length: 450, dtype: float64

1) SGD Classifier:



SDG

SGDClassifier(max\_iter=450, tol=0.01)

Training score: 0.07407407407407407

[[4 0 0 ... 0 0 0]

[0 0 0 ... 1 0 0]

[0 0 0 ... 0 0 0]

...

[0 0 0 ... 0 0 0]

[0 0 0 ... 0 0 0]

[0 0 0 ... 0 0 0]]

precision recall f1-score support

0.0 1.00 1.00 1.00 4

10.0 0.00 0.00 0.00 1

15.0 0.00 0.00 0.00 1

20.0 0.00 0.00 0.00 1

50.0 0.00 0.00 0.00 1

60.0 0.00 0.00 0.00 2

90.0 0.00 0.00 0.00 1

110.0 0.00 0.00 0.00 1

150.0 0.00 0.00 0.00 2

170.0 0.00 0.00 0.00 0

180.0 0.00 0.00 0.00 2

190.0 0.00 0.00 0.00 2

210.0 0.00 0.00 0.00 1

240.0 0.00 0.00 0.00 2

250.0 0.00 0.00 0.00 1

260.0 0.00 0.00 0.00 1

280.0 0.00 0.00 0.00 1

300.0 0.00 0.00 0.00 1

310.0 0.00 0.00 0.00 1

320.0 0.00 0.00 0.00 1

340.0 0.00 0.00 0.00 2

380.0 0.00 0.00 0.00 1

460.0 0.00 0.00 0.00 1

500.0 0.00 0.00 0.00 0

520.0 0.00 0.00 0.00 1

530.0 0.00 0.00 0.00 1

540.0 0.00 0.00 0.00 1

630.0 0.00 0.00 0.00 1

660.0 0.00 0.00 0.00 1

670.0 0.00 0.00 0.00 1

720.0 0.00 0.00 0.00 1

740.0 0.00 0.00 0.00 2

770.0 0.00 0.00 0.00 2

790.0 0.00 0.00 0.00 1

900.0 0.00 0.00 0.00 0

980.0 0.00 0.00 0.00 1

1110.0 0.00 0.00 0.00 1

accuracy 0.09 45

macro avg 0.03 0.03 0.03 45

weighted avg 0.09 0.09 0.09 45

SGDClassifier(max\_iter=450, tol=0.01)

Score: 0.9703703703703703

[[8 0 0]

[0 3 0]

[0 0 4]]

precision recall f1-score support

0 1.00 1.00 1.00 8

1 1.00 1.00 1.00 3

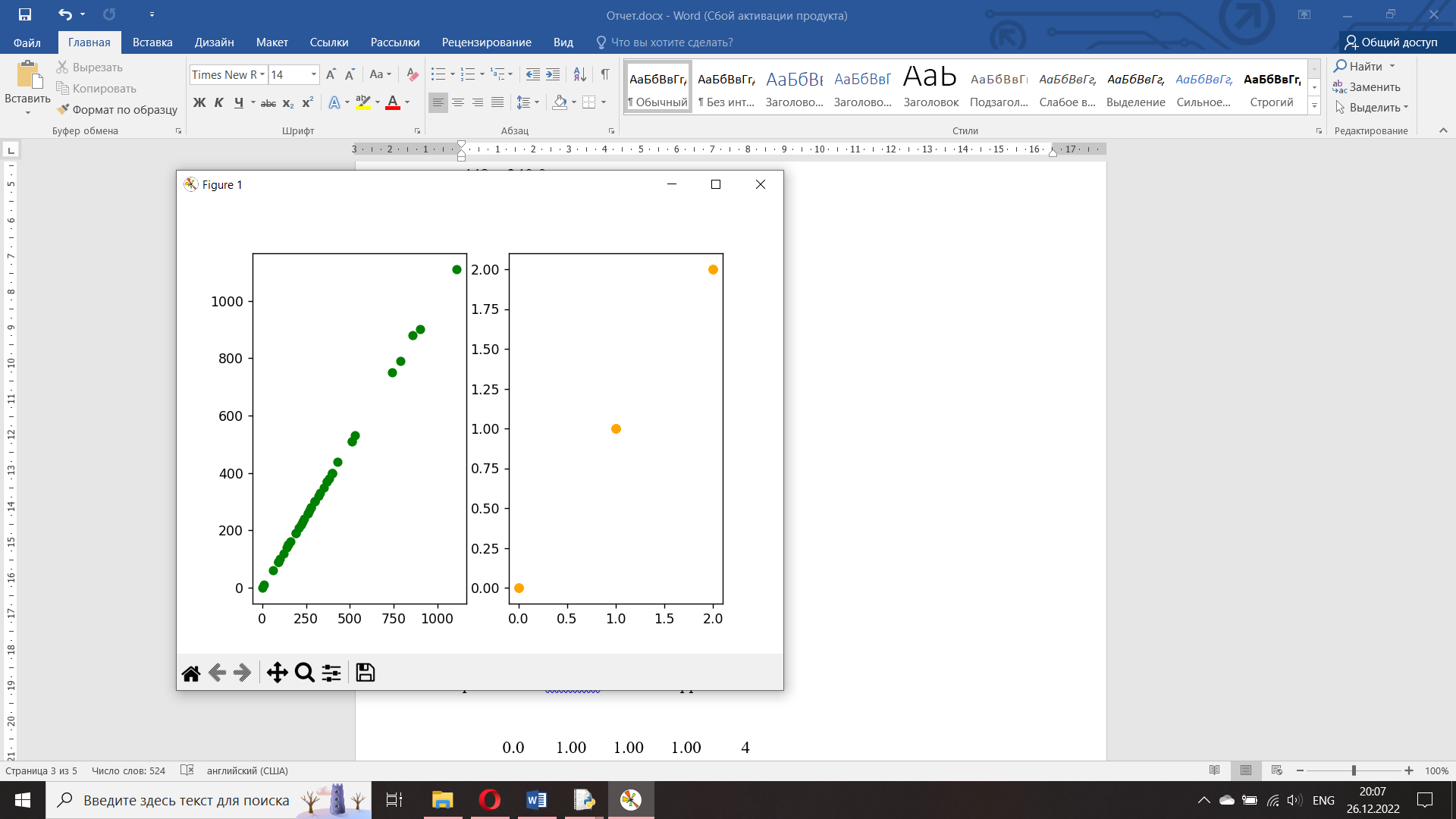
2 1.00 1.00 1.00 4

accuracy 1.00 15

macro avg 1.00 1.00 1.00 15

weighted avg 1.00 1.00 1.00 15

2) Support Vector Machines:



SVM

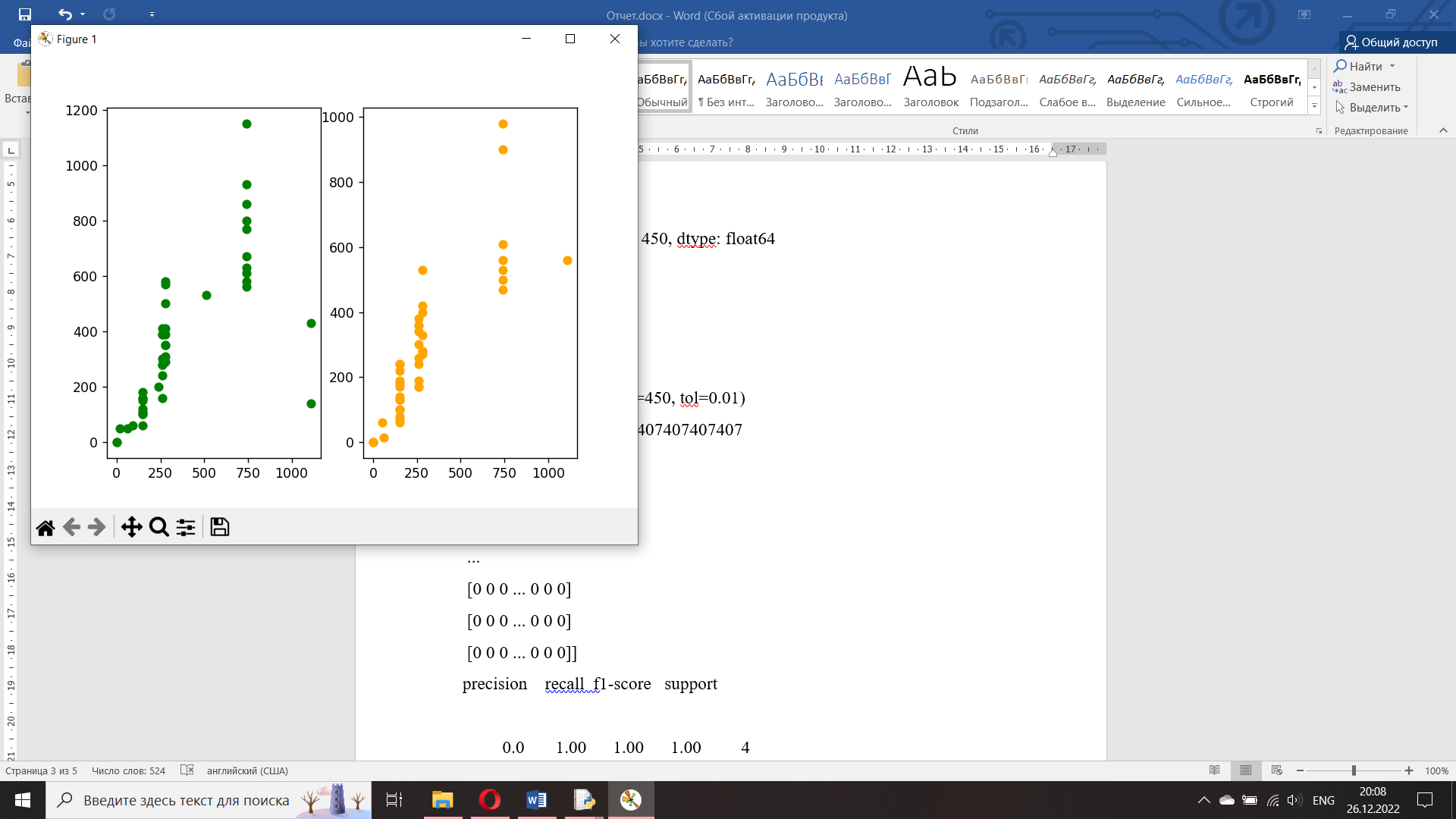
accuracy of the model 1:

0.9333333333333333

accuracy of the model 2:

1.0

3) Logistic Regression:



Logistic Regression

score 1:

0.18271604938271604

precision recall f1-score support

0.0 1.00 1.00 1.00 2

20.0 0.00 0.00 0.00 0

50.0 0.00 0.00 0.00 2

60.0 0.00 0.00 0.00 2

90.0 0.00 0.00 0.00 0

100.0 0.00 0.00 0.00 2

110.0 0.00 0.00 0.00 1

120.0 0.00 0.00 0.00 1

140.0 0.00 0.00 0.00 1

150.0 0.11 1.00 0.20 1

160.0 0.00 0.00 0.00 3

180.0 0.00 0.00 0.00 1

200.0 0.00 0.00 0.00 1

240.0 0.00 0.00 0.00 1

260.0 0.00 0.00 0.00 0

280.0 0.00 0.00 0.00 2

290.0 0.00 0.00 0.00 1

300.0 0.00 0.00 0.00 1

310.0 0.00 0.00 0.00 1

350.0 0.00 0.00 0.00 2

390.0 0.00 0.00 0.00 3

410.0 0.00 0.00 0.00 2

430.0 0.00 0.00 0.00 1

500.0 0.00 0.00 0.00 1

510.0 0.00 0.00 0.00 0

530.0 0.00 0.00 0.00 1

560.0 0.00 0.00 0.00 1

570.0 0.00 0.00 0.00 1

580.0 0.00 0.00 0.00 2

610.0 0.00 0.00 0.00 1

630.0 0.00 0.00 0.00 1

670.0 0.00 0.00 0.00 1

740.0 0.00 0.00 0.00 0

770.0 0.00 0.00 0.00 1

800.0 0.00 0.00 0.00 1

860.0 0.00 0.00 0.00 1

930.0 0.00 0.00 0.00 1

1110.0 0.00 0.00 0.00 0

1150.0 0.00 0.00 0.00 1

accuracy 0.07 45

macro avg 0.03 0.05 0.03 45

weighted avg 0.05 0.07 0.05 45

score 2:

0.1580246913580247

precision recall f1-score support

0.0 1.00 1.00 1.00 3

15.0 0.00 0.00 0.00 1

50.0 0.00 0.00 0.00 0

60.0 0.00 0.00 0.00 2

70.0 0.00 0.00 0.00 1

80.0 0.00 0.00 0.00 1

100.0 0.00 0.00 0.00 2

130.0 0.00 0.00 0.00 2

140.0 0.00 0.00 0.00 1

150.0 0.00 0.00 0.00 0

170.0 0.00 0.00 0.00 3

180.0 0.00 0.00 0.00 2

190.0 0.00 0.00 0.00 2

220.0 0.00 0.00 0.00 1

240.0 0.00 0.00 0.00 3

260.0 0.10 1.00 0.18 1

270.0 0.00 0.00 0.00 1

280.0 0.29 1.00 0.44 2

300.0 0.00 0.00 0.00 1

330.0 0.00 0.00 0.00 1

340.0 0.00 0.00 0.00 1

360.0 0.00 0.00 0.00 2

380.0 0.00 0.00 0.00 1

400.0 0.00 0.00 0.00 1

420.0 0.00 0.00 0.00 1

470.0 0.00 0.00 0.00 1

500.0 0.00 0.00 0.00 1

530.0 0.00 0.00 0.00 2

560.0 0.00 0.00 0.00 2

610.0 0.00 0.00 0.00 1

740.0 0.00 0.00 0.00 0

900.0 0.00 0.00 0.00 1

980.0 0.00 0.00 0.00 1

1110.0 0.00 0.00 0.00 0

accuracy 0.13 45

macro avg 0.04 0.09 0.05 45

weighted avg 0.08 0.13 0.09 45

**Вывод:**

SVM пытается максимизировать разницу между ближайшими векторами поддержки, а LR - вероятностью апостериорного класса. Таким образом, SVM находит решение, которое является максимально возможным для обеих категорий, в то время как LR не обладает этим свойством. LR более чувствителен к выбросам, чем SVM, потому что функция стоимости LR расходится быстрее, чем функция SVM. SDG и SVM всегда используют полные данные и решают задачу выпуклой оптимизации относительно этих точек данных. SDG прост в реализации, но чувствителен к масштабированию функций.