МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет информационных систем и технологий
Кафедра «Измерительно-вычислительные комплексы»

Отчет по лабораторной работе №5
по дисциплине «Методы искусственного интеллекта»

Выполнила

ст. гр. ИСТбд-41 Карташова М.В.

Проверил:

Шишкин В.В.

Ульяновск, 2022

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
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.

Ход выполнения работы:

1. После ознакомления, для лабораторной работы были выбраны датасет Loan_Data.csv и классификаторы: DecisionTreeClassifier, KNeighborsClassifier, RandomForestClassifier, GradientBoostingClassifier. Данный выбор утвержден у преподавателя.
2. Для классификаторов определен целевой столбец – Loan_Status (Статус одобрения кредита). Для определения набора признаков, которые сильнее всего влияют на значение целевого столбца, был написан скрипт, выводящий столбчатую диаграмму важности всех признаков. Результат, представлен на рис.1.

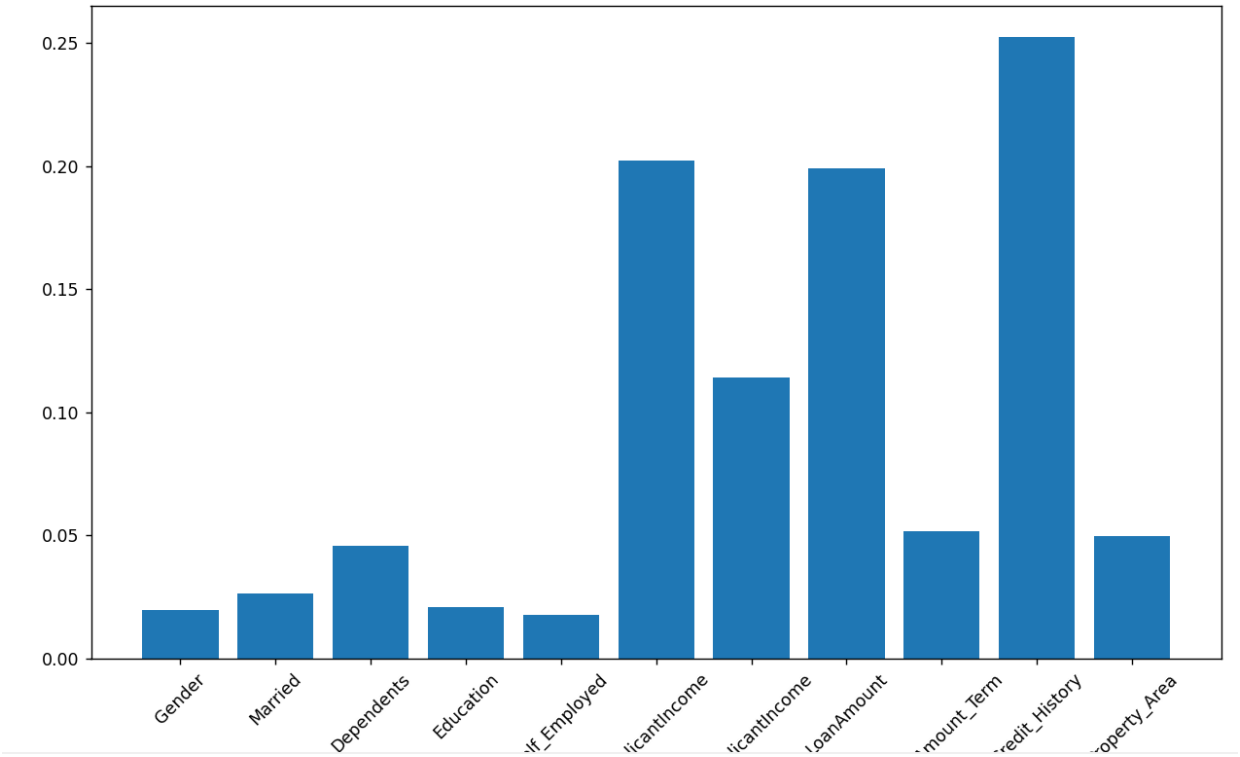

Рис.1. Оценка влияния признаков

Таким образом, больше всего влияния на целевой столбец оказывают признаки: AmpplicantIncome (Доход заявителя), LoanAmount (Сумма кредита) и Credit_History (Кредитная история).

Новый вид датасета, только с главными признаками и целевым столбцом можно увидеть на рис.2.

Размерность датасета : (614, 4)

	ApplicantIncome	LoanAmount	Credit_History	Loan_Status
0	5849	NaN	1.0	Y
1	4583	128.0	1.0	N
2	3000	66.0	1.0	Y
3	2583	120.0	1.0	Y
4	6000	141.0	1.0	Y
..	...	...	...	...
609	2900	71.0	1.0	Y
610	4106	40.0	1.0	Y
611	8072	253.0	1.0	Y
612	7583	187.0	1.0	Y
613	4583	133.0	0.0	N

Рис.2. Датасет для обучения и классификации

3. Далее были подготовлены данные к обучению и пробовалась провести обучение моделей на сырых данных. Так как в сырых данных оказались пропуски компилятор выдавал ошибку.

```
ValueError: Input X contains NaN.
```

Рис.3. Ошибка компиляции

```
Кол-во пропусков изначально:
 ApplicantIncome     0
LoanAmount          22
Credit_History      50
Loan_Status          0
```

Рис.4. Кол-во пропусков в датасете до очистки

4. Затем данные были почищены путем замены пропусков в одном столбце на значение медианы, остальные строки с пропусками были удалены. Проведено обучение моделей на очищенных данных.
5. Далее была произведена оценка моделей и результаты представлены в табличной и графической форме.
    - DecisionTreeClassifier:

```
Отчет, показывающий основные метрики классификации методом <Дерево решений>
              precision    recall  f1-score   support

           N       0.56      0.43      0.49        42
           Y       0.78      0.86      0.82        99

    accuracy                           0.73       141
   macro avg       0.67      0.64      0.65       141
weighted avg       0.72      0.73      0.72       141
```

Рис.5. Оценка DecisionTreeClassifier в табличной форме

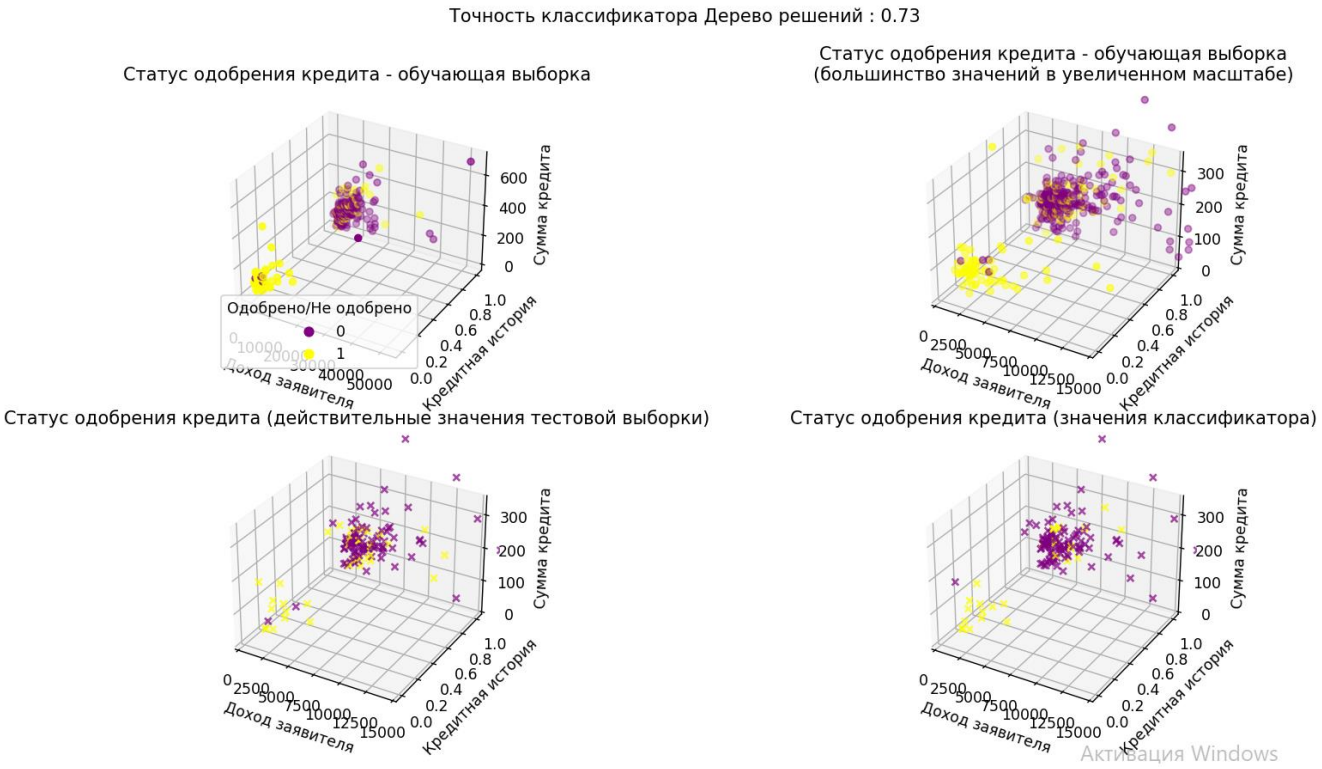

Рис.6. Оценка DecisionTreeClassifier в графической форме

По нижним двум графикам можно наглядно увидеть, сравнив, какие именно точки классификатор определил правильно, а которые не совпадают с действительным значением. А смотря на верхние графики можно понять, на какие группы делятся данные в общем и где эти группы представлены в 3-х мерном пространстве признаков.

- GradientBoostingClassifier

```
Отчет, показывающий основные метрики классификации методом <Повышение градиента>
              precision    recall  f1-score   support

           N       0.73      0.38      0.50        42
           Y       0.78      0.94      0.85        99

    accuracy                           0.77       141
   macro avg       0.75      0.66      0.68       141
weighted avg       0.77      0.77      0.75       141


Process finished with exit code 0
```

Рис.7. Оценка GradientBoostingClassifier в табличной форме

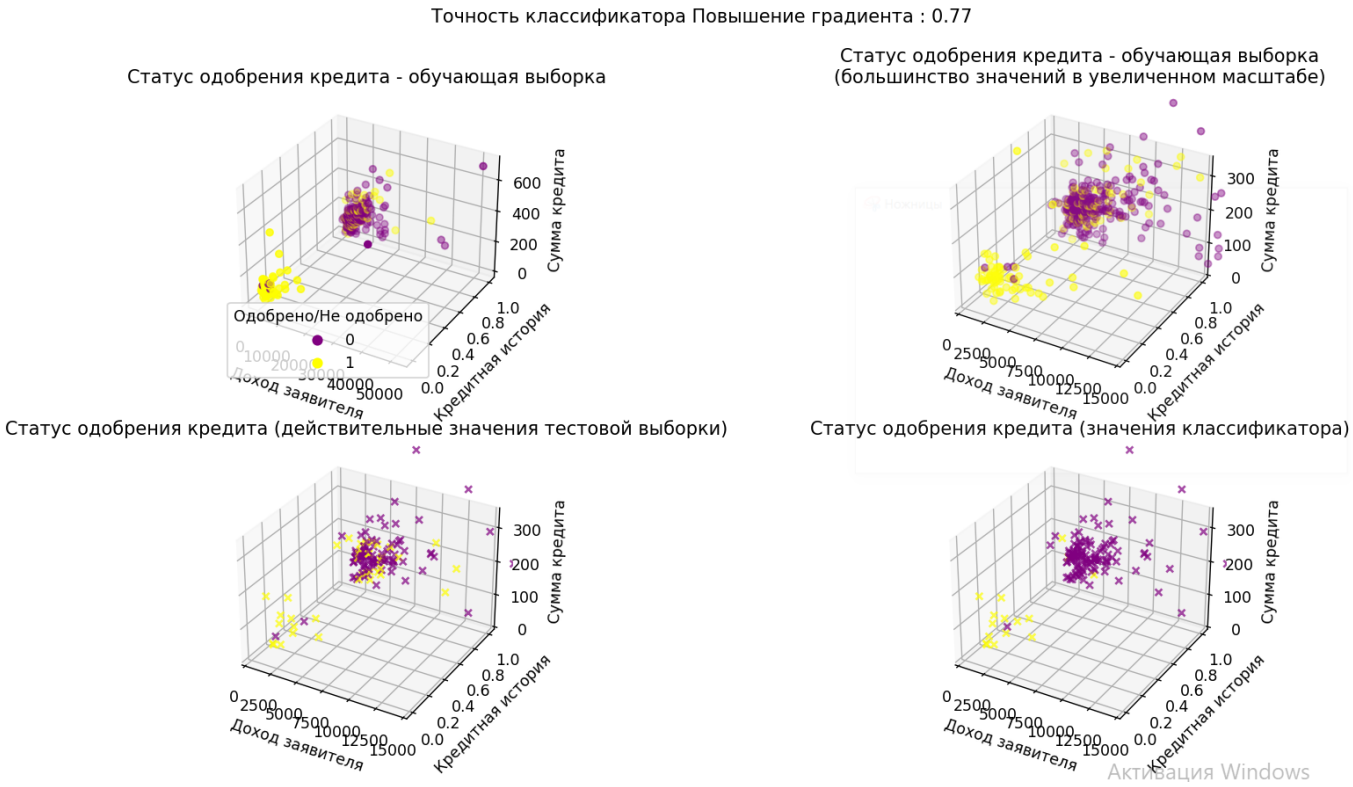

Рис.8. Оценка GradientBoostingClassifier в графической форме

- RandomForestClassifier:

Отчет, показывающий основные метрики классификации методом <Случайный лес>

	precision	recall	f1-score	support
N	0.62	0.43	0.51	42
Y	0.79	0.89	0.83	99
accuracy			0.75	141
macro avg	0.70	0.66	0.67	141
weighted avg	0.74	0.75	0.74	141

Рис.9. Оценка RandomForestClassifier в табличной форме

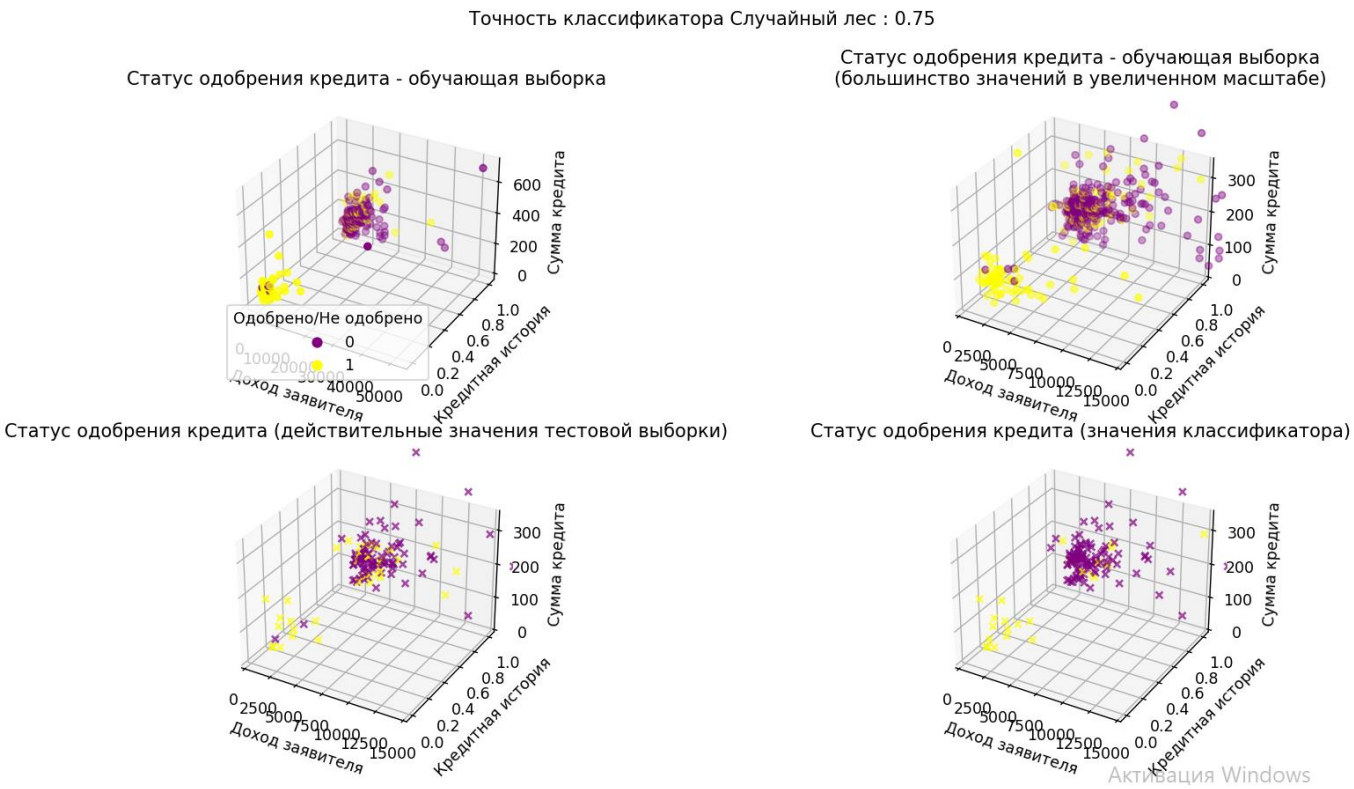

Рис.10. Оценка RandomForestClassifier в графической форме

- KNeighborsClassifier

Отчет, показывающий основные метрики классификации методом <К ближайших соседей>

	precision	recall	f1-score	support
N	0.41	0.31	0.35	42
Y	0.73	0.81	0.77	99
accuracy			0.66	141
macro avg	0.57	0.56	0.56	141
weighted avg	0.64	0.66	0.64	141

Рис.11. Оценка KNeighborsClassifier в табличной форме

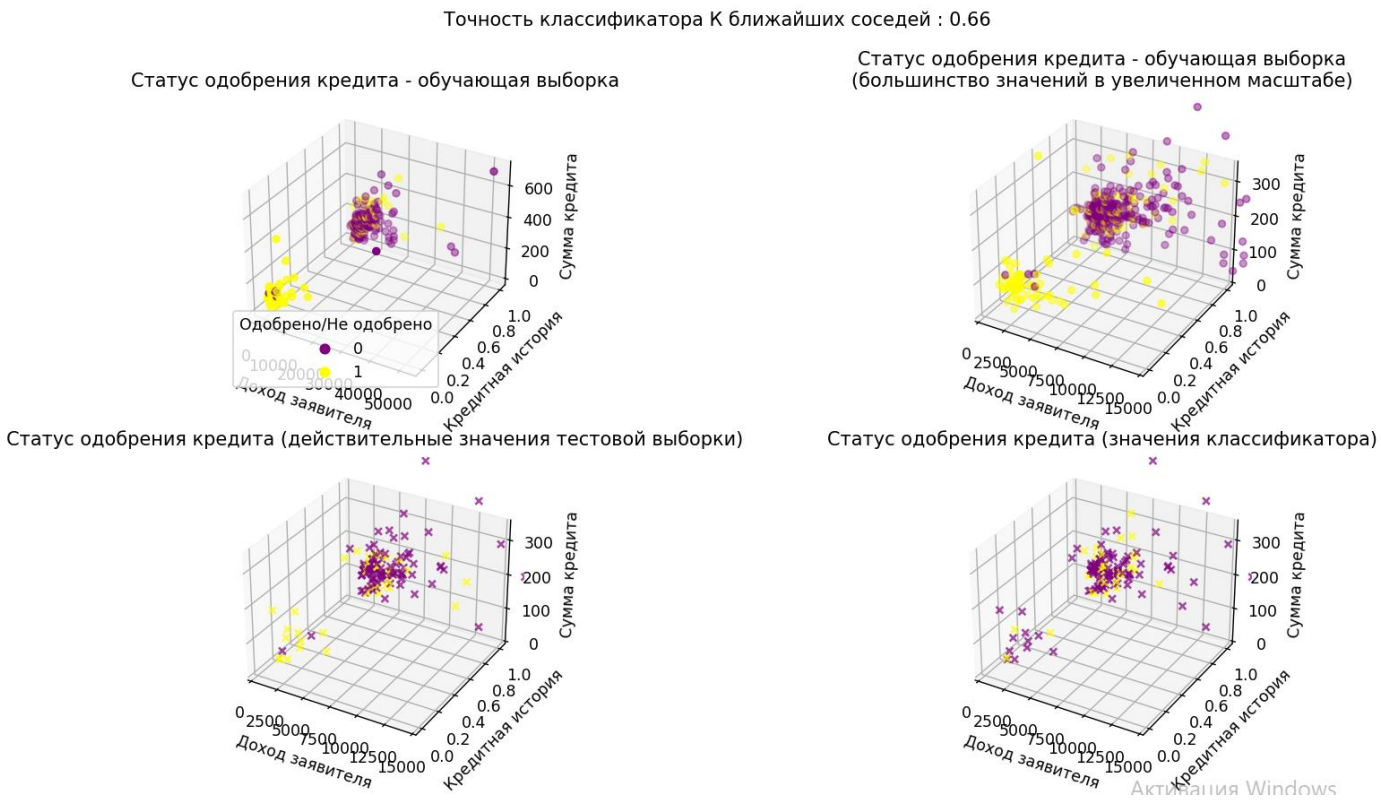

Рис.12. Оценка KNeighborsClassifier в графической форме

Вывод: в ходе данной лабораторной работы, я познакомилась с несколькими классификаторами sklearn, научилась выбирать главные признаки в датасете для классификации, готовить и очищать данные для классификации. Проведя оценку работ классификаторов было выявлено, что на данном наборе данных лучше всего с классификацией данных справился GradientBoostingClassifier, а хуже всего KNeighborsClassifier.