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»

Исследование инструментов классификации библиотеки Scikit-learn

Отчёт по лабораторной работе №5

Вариант №12

Выполнила:
студентка группы ИСТбд-42
Кучина Анна
Проверил:
доцент кафедры ИВК, к.т.н.
Шишкин В. В.

Ульяновск
2022

Задание на лабораторную работу:

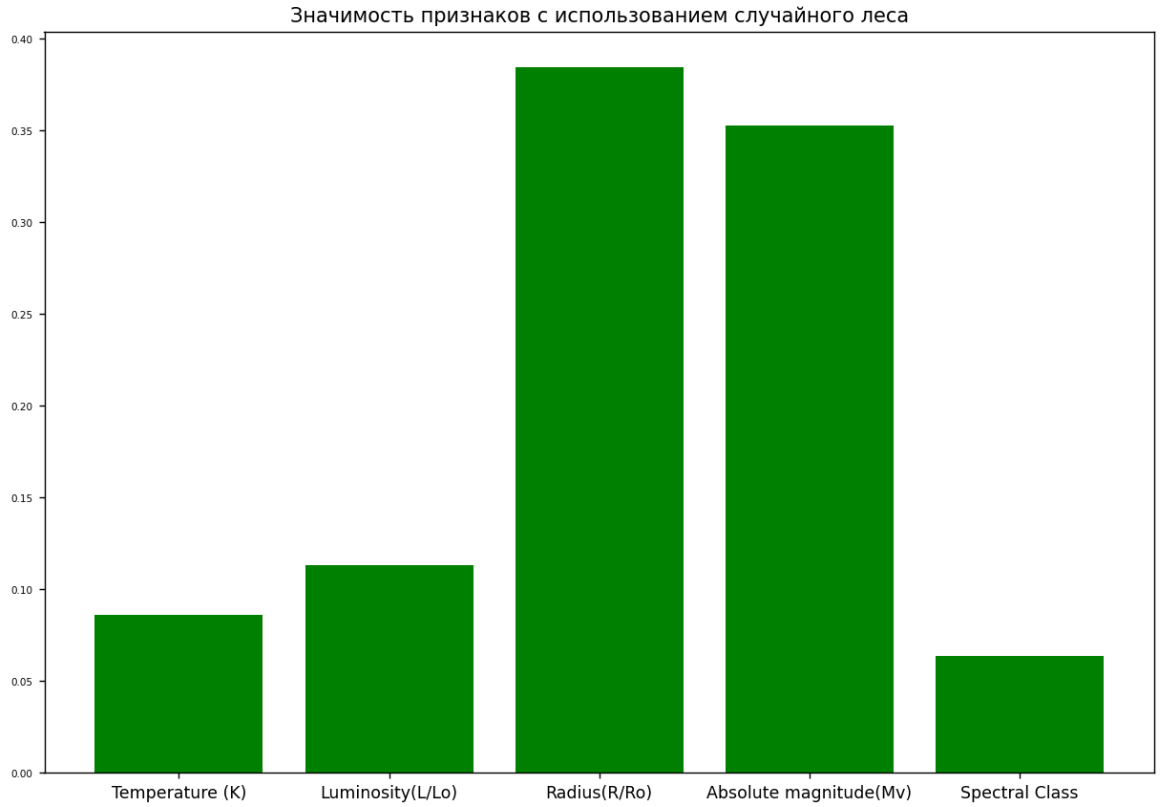
1. Ознакомиться с классификаторами библиотеки Scikit-learn
2. Выбрать для исследования не менее 3 классификаторов
3. Выбрать набор данных для задач классификации из открытых источников
4. Выбор классификаторов и набора данных утвердить у преподавателя (не должно быть полного совпадения с выбором другого 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.

Выполнение работы:

1. Для исследования были выбраны следующие классификаторы:
    a. К ближайших соседей (knn)
    b. Наивный Байесовский классификатор (GaussianNB)
    c. Случайный лес (Random Forest)

2. Для обучения модели был выбран набор данных со следующими столбцами:
   Absolute Temperature (in K) – абсолютная температура в Кельвинах
   Relative Luminosity (L/Lo) – относительная светимость
   Relative Radius (R/Ro) – относительный радиус
   Absolute Magnitude (Mv) – абсолютная величина
   Spectral Class (O,B,A,F,G,K,,M) – спектральный класс
   Star Type **(Красный карлик = 0, Коричневый карлик = 1, Белый карлик = 2, Главная последовательность = 3, Гигант = 4, Сверхгигант = 5)**
   Lo = 3.828 x 10^26 Watts (Avg Luminosity of Sun)
   Ro = 6.9551 x 10^8 m (Avg Radius of Sun)
   В датасете 240 строк
   https://www.kaggle.com/datasets/deepu1109/star-dataset?resource=download

3. За целевой столбец для каждого классификатора был выбран Тип звезды (Star Type), поскольку создателем набора данных он подразумевался как выходной. К тому же все прочие столбцы представляют собой критерии для определения типа звезды и ее положение на диаграмме Герцшпрунга-Рассела. Как правило, тип звезды определяется по ее температуре, размеру и светимости. Однако после проведения анализа критериев с помощью алгоритма случайного леса был сделан вывод о том, что наиболее значимыми критериями являются радиус и абсолютная величина звезды



4. При подготовке данных к обучению было выявлено, что данный набор данных не имеет пустых значений, а значит, не нуждается в дополнительной очистке:

```
RangeIndex: 240 entries, 0 to 239
Data columns (total 5 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Temperature (K)        240 non-null    int64
 1   Luminosity(L/Lo)       240 non-null    float64
 2   Radius(R/Ro)           240 non-null    float64
 3   Absolute magnitude(Mv) 240 non-null    float64
 4   Spectral Class         240 non-null    int64
```

Однако для работы с алгоритмами библиотеки sklearn данные столбца Спектральный класс были переведены в эквивалентные числовые значения от 0 до 6:

O = 0, B = 1, A = 2, F = 3, G = 4, K = 5, N = 6

Всего в тестовой выборке 96 элементов для классификации

5. Результат работы модели, обученной на трех классификаторах:

К ближайших соседей

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	0.79	0.69	0.73	16
2	0.76	0.84	0.80	19
3	0.83	0.88	0.86	17
4	0.62	0.68	0.65	19
5	0.50	0.36	0.42	11
accuracy			0.76	96
macro avg	0.75	0.74	0.74	96
weighted avg	0.75	0.76	0.76	96

Точность алгоритма - 76%

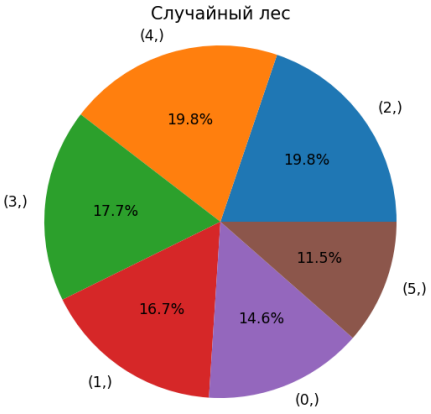
Наивный Байесовский классификатор

	precision	recall	f1-score	support
0	0.88	1.00	0.93	14
1	0.42	0.94	0.58	16
2	0.00	0.00	0.00	19
3	0.93	0.82	0.87	17
4	1.00	0.95	0.97	19
5	1.00	1.00	1.00	11
accuracy			0.75	96
macro avg	0.70	0.78	0.73	96
weighted avg	0.67	0.75	0.69	96

Точность алгоритма – 75%

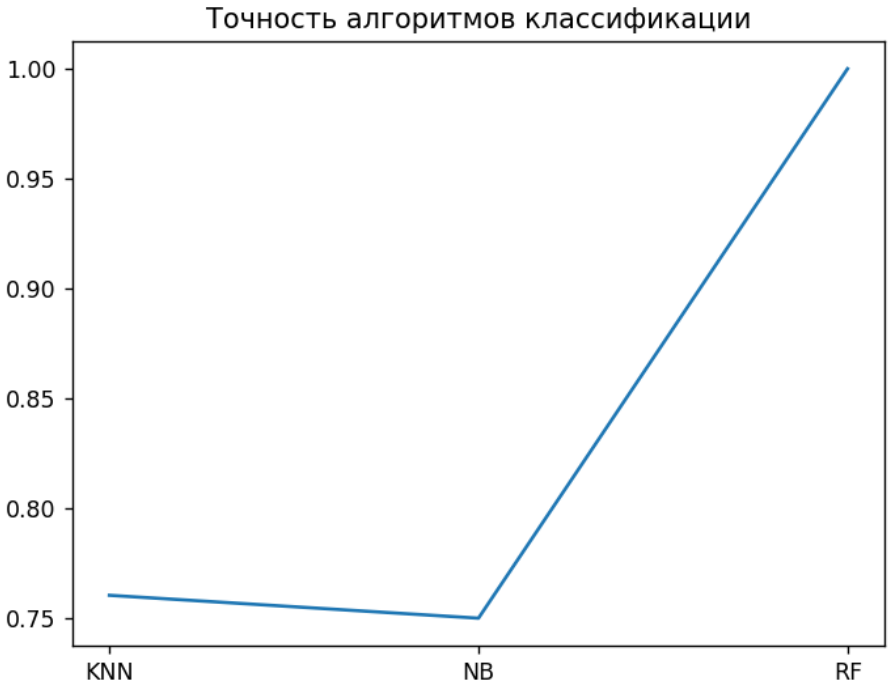
Случайный лес

	precision	recall	f1-score	support
0	1.00	1.00	1.00	14
1	1.00	1.00	1.00	16
2	1.00	1.00	1.00	19
3	1.00	1.00	1.00	17
4	1.00	1.00	1.00	19
5	1.00	1.00	1.00	11
accuracy			1.00	96
macro avg	1.00	1.00	1.00	96
weighted avg	1.00	1.00	1.00	96



Точность алгоритма – 100%

6. При анализе полученных данных были получены следующие оценки точности



Вывод по работе: В ходе работы были изучены три классификатора библиотеки sklearn. На выбранном наборе данных точнее всего был алгоритм классификации Случайный лес. Наименьшей точностью обладает Наивный Байесовский классификатор.