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:
   1. К ближайших соседей (knn)
   2. Наивный Байесовский классификатор (GaussianNB)
   3. 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

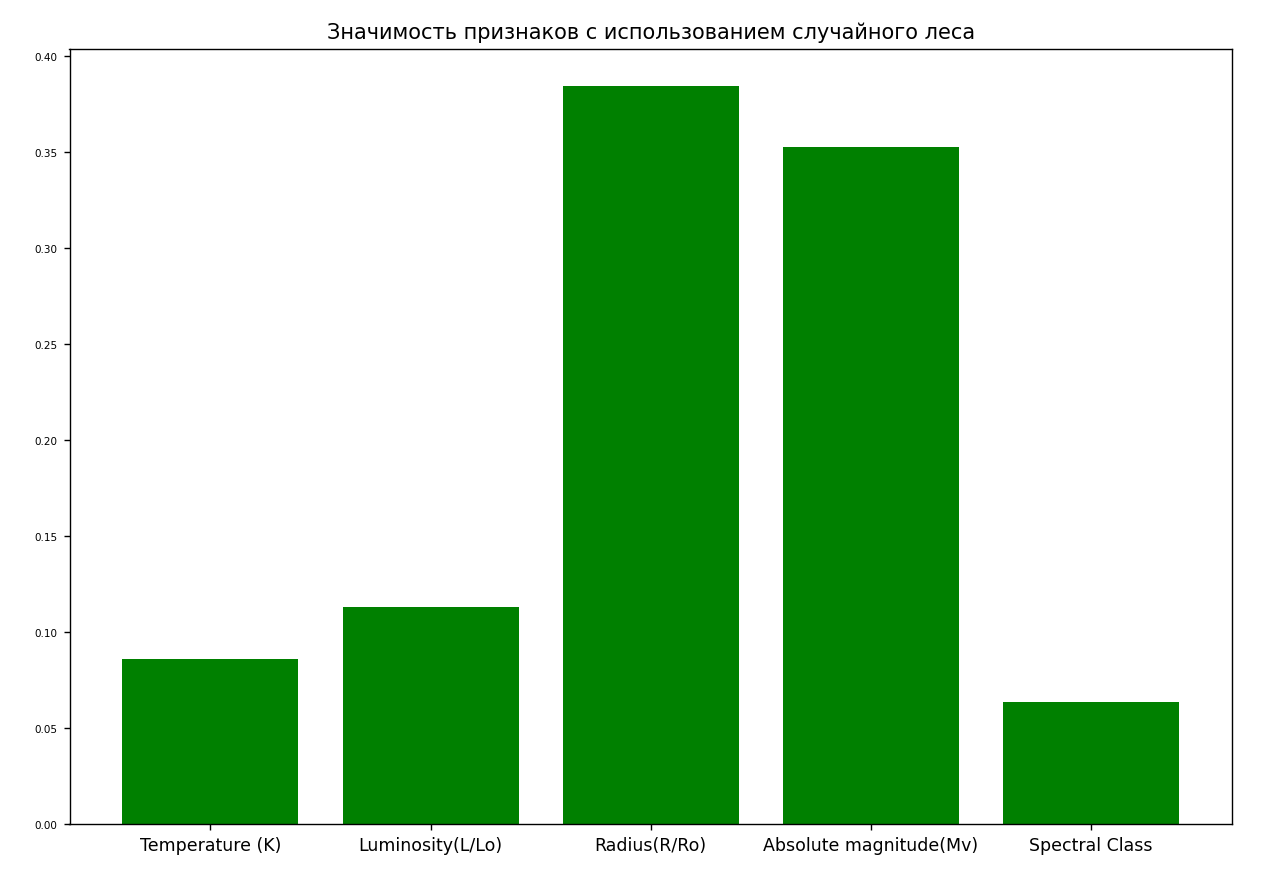
Spectral Class (O,B,A,F,G,K,,M) – спектральный класс

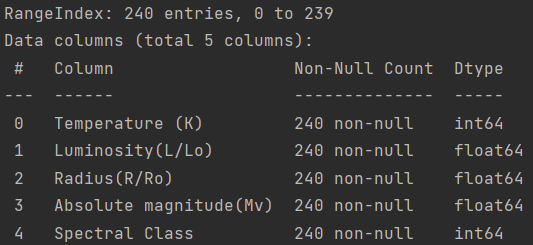
Star Type \*\*(Красный карлик = 0, Коричневый карлик = 1, Белый карлик = 2, Главная последовательность = 3, Гигант = 4, Сверхгигант = 5)\*\*

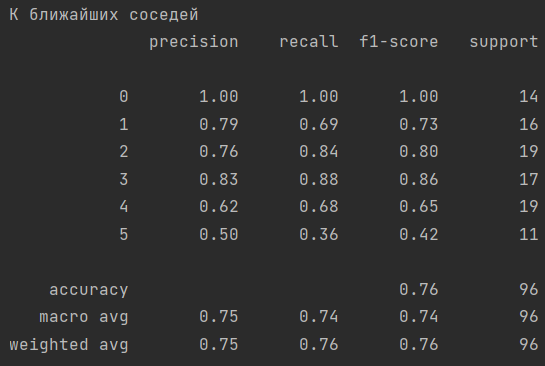
Lo = 3.828 x 10^26 Watts (Avg Luminosity of Sun)

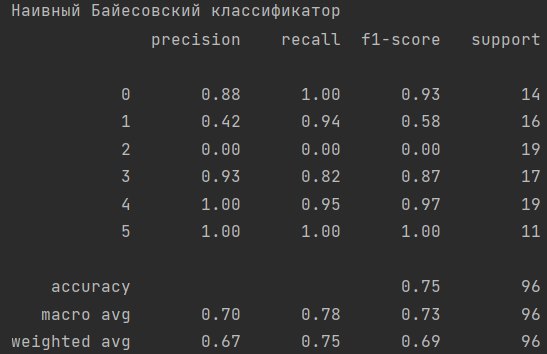
Ro = 6.9551 x 10^8 m (Avg Radius of Sun)  
В датасете 240 строк  
https://www.kaggle.com/datasets/deepu1109/star-dataset?resource=download

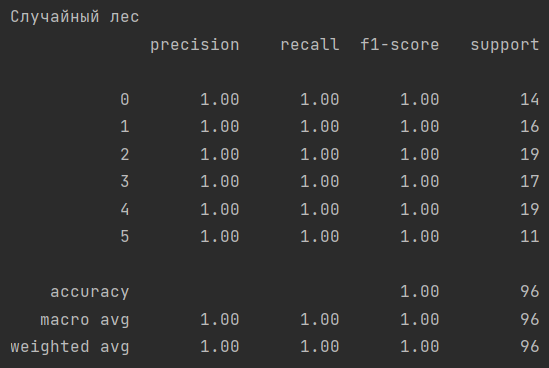
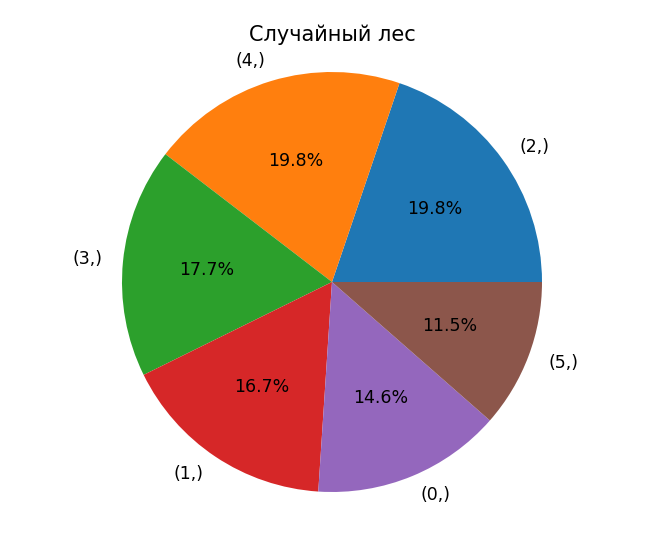
1. За целевой столбец для каждого классификатора был выбран Тип звезды (Star Type), поскольку создателем набора данных он подразумевался как выходной. К тому же все прочие столбцы представляют собой критерии для определения типа звезды и ее положение на диаграмме Герцшпрунга-Рассела. Как правило, тип 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



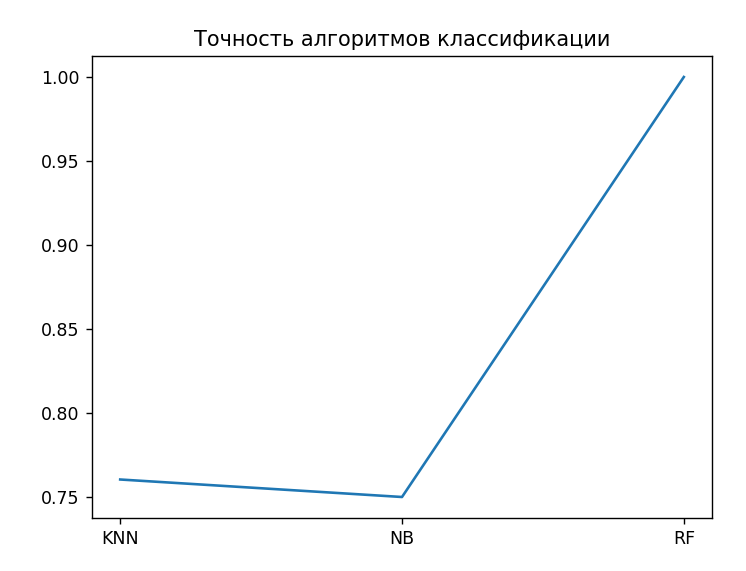
1. При подготовке данных к обучению было выявлено, что данный набор данных не имеет пустых значений, а значит, не нуждается в дополнительной очистке:  
     
   Однако для работы с алгоритмами библиотеки sklearn данные столбца Спектральный класс были переведены в эквивалентные числовые значения от 0 до 6:  
   O = 0, B = 1, A = 2, F = 3, G = 4, K = 5, N = 6  
   Всего в тестовой выборке 96 элементов для классификации
2. Результат работы модели, обученной на трех классификаторах:

Точность алгоритма - 76%  
 

Точность алгоритма – 75%  
 

Точность алгоритма – 100%

1. 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.