МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И.УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЁТ

по лабораторной работе №7
по дисциплине «машинное обучение»
Тема: классификация (байесовские методы, деревья)

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

Цель работы

Ознакомиться с методами классификации модуля Sklearn.

1. Байесовские методы

1. Произведена загрузка данных. Часть набора данных представлена на листинге 1.

Листинг 1. Набор данных

```
      0    1    2    3               4
0    5.1  3.5  1.4  0.2     Iris-setosa
1    4.9  3.0  1.4  0.2     Iris-setosa
2    4.7  3.2  1.3  0.2     Iris-setosa
3    4.6  3.1  1.5  0.2     Iris-setosa
4    5.0  3.6  1.4  0.2     Iris-setosa
..   ...  ...  ...  ...             ...
145  6.7  3.0  5.2  2.3  Iris-virginica
146  6.3  2.5  5.0  1.9  Iris-virginica
147  6.5  3.0  5.2  2.0  Iris-virginica
148  6.2  3.4  5.4  2.3  Iris-virginica
149  5.9  3.0  5.1  1.8  Iris-virginica

[150 rows x 5 columns]
```

2. Произведена классификация наблюдений наивным байесовским классификатором. Количество наблюдений, которые были неправильно определены, и точность классификации представлены на листинге 2.

Листинг 2. Классификация наивным байесовским методом

```
Количество ошибок: 6
score: 0.92
```

3. Атрибуты `GaussianNB`:
   - `class_count_` — количество элементов в каждом классе;
   - `class_prior_` — априорные вероятности каждого класса;
   - `classes_` — метки классов;
   - `epsilon_` — сумма дисперсий;
   - `sigma_` — дисперсия каждого признака в каждом классе;
   - `theta_` — среднее каждого признака по каждому классу.

4. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. Результат на рис. 1.

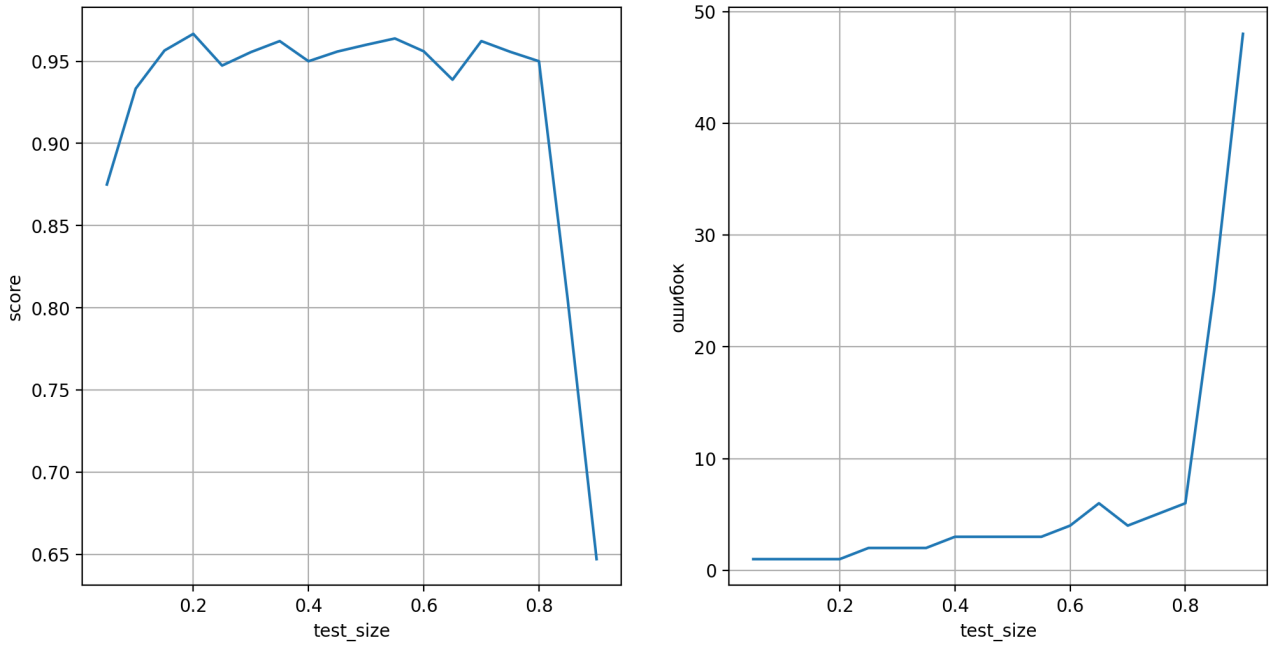

Рисунок 1 – График зависимости точности и количества ошибок от `test_size`

Как видно, существенное понижение результатов получается только при небольшом объеме выборки для обучения. Таким образом, в данном случае наивному байесовскому классификатору достаточно небольшого количества данных для обучения.

5. Аналогичные действия проведены для `MultinomialNB`, `ComplementNB`, `Bernoull` Результаты на рис. 2.

- `MultinomialNB` — реализует мультиномиальный наивный байесовский классификатор. Можно использовать, если признаки взяты из мультиномиального распределения, например, для частоты слов в тексте.
- `CompletementNB` — может применятся в тех же случаях, что и `MNB`. В отличие от `MNB`, здесь считается вероятность, что образец не принадлежит к какому-либо классу, и берется тот класс, для которого наименьшая вероятность, что образец к нему не принадлежит. В некоторых случаях этот подход может быть более стабилен.
- `BernoulliNB` — используется, если признаки являются бинарными.

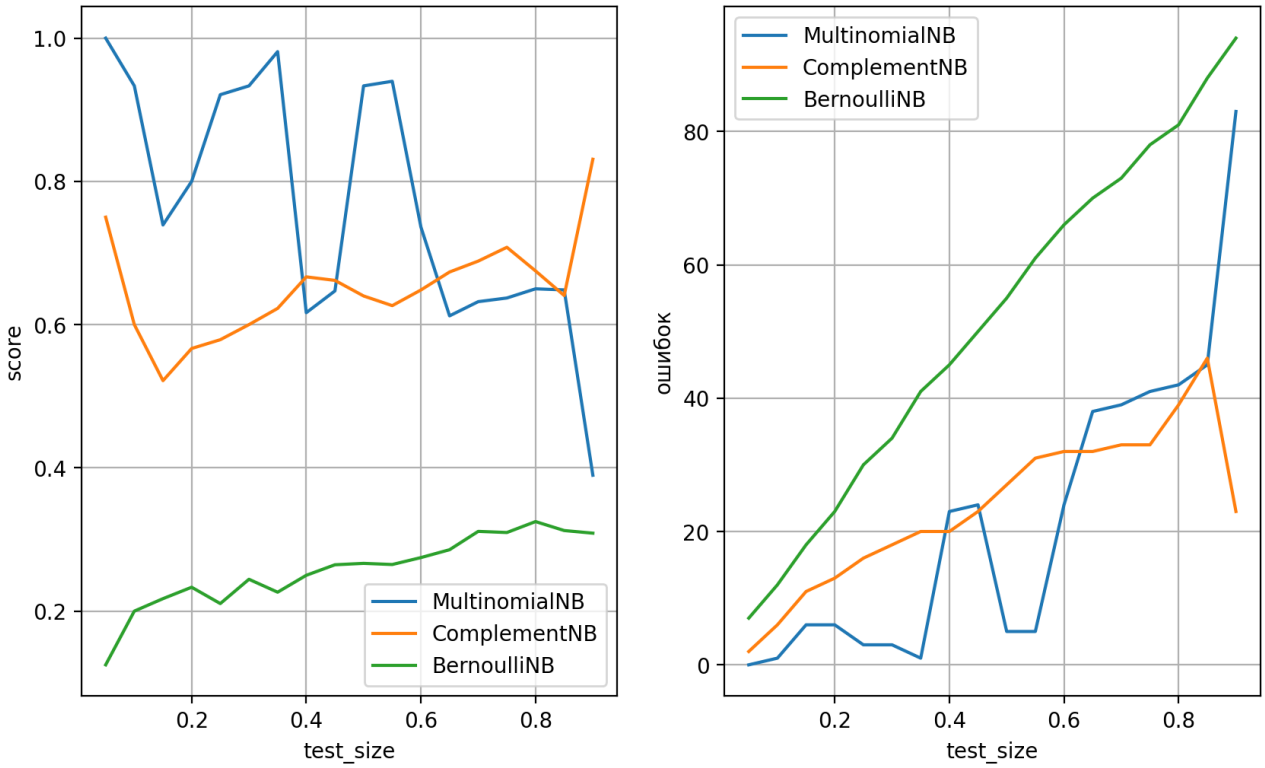

Рисунок 2 – Различные методы

2. Деревья решений

1. Произведена классификация с помощью дерева решений на тех же данных. Количество ошибок и `score` представлены на листинге 3.

Листинг 3. Результаты

```
Количество ошибок: 5
score: 0.9333333333333333
```

2. Количество листьев и глубина представлены на листинге 4

Листинг 4. Количество листьев и глубина

```
num_leaves: 6
depth: 5
```

3. Полученное дерево представлено на рис. 3.

   Во всех элементах дерева, кроме листьев, в первой строчке написано условие ветвления. Оранжевые листья означают, что образец принадлежит к 0-му классу, зеленые — к 1-му, синие — ко 2-му.

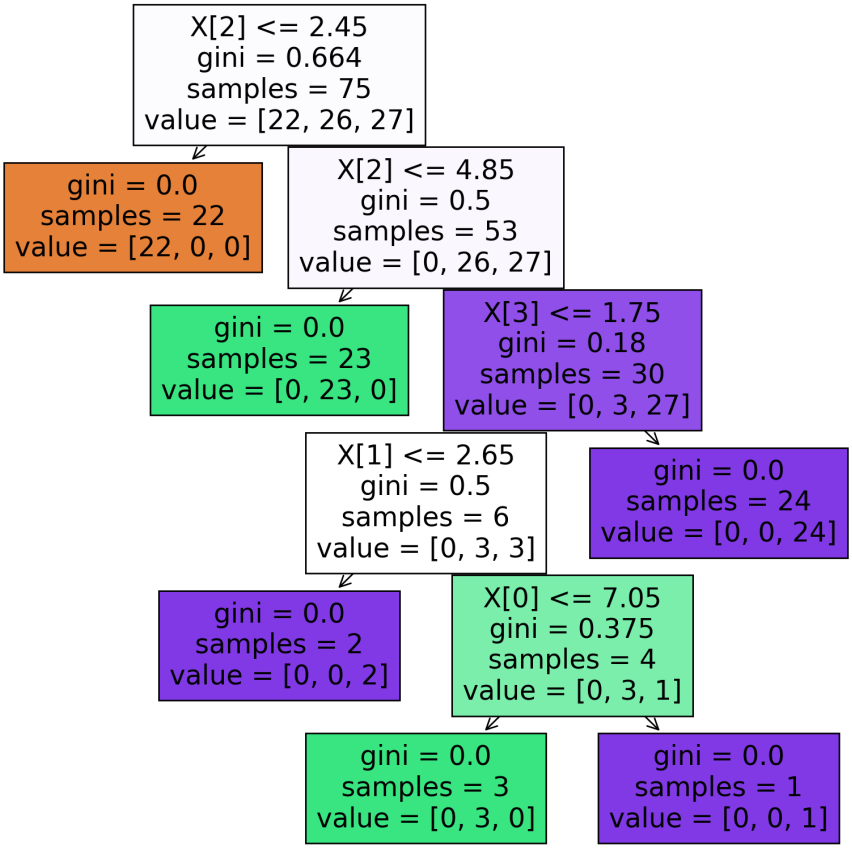

Рисунок 3 – Дерево

4. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. Результат на рис. 4.

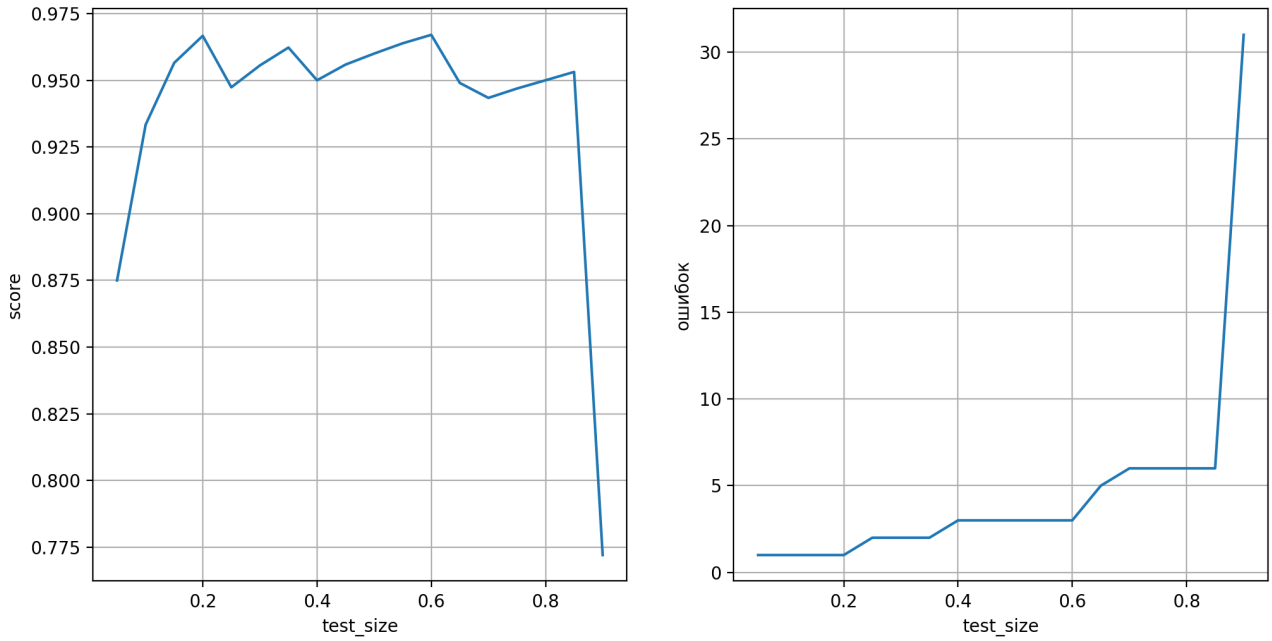

Рисунок 4 – Зависимость точности от размера тестовой выборки

Как и в предыдущем случае, классификация становится намного хуже только при небольших значениях набора для обучения. Количество ошибок возрастает скачкообразно по мере уменьшения обучающего набора.

5. Исследована работа классификатора при разных значениях параметров. Результаты на рис. 5–9.

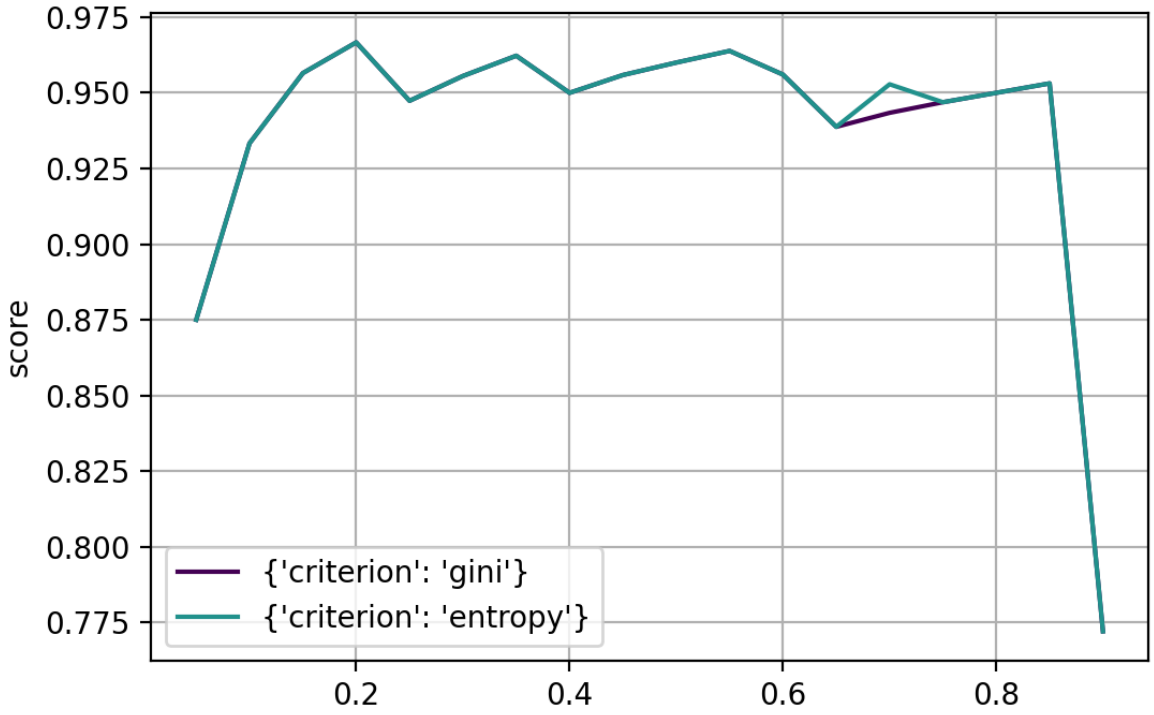

Рисунок 5 – criterion

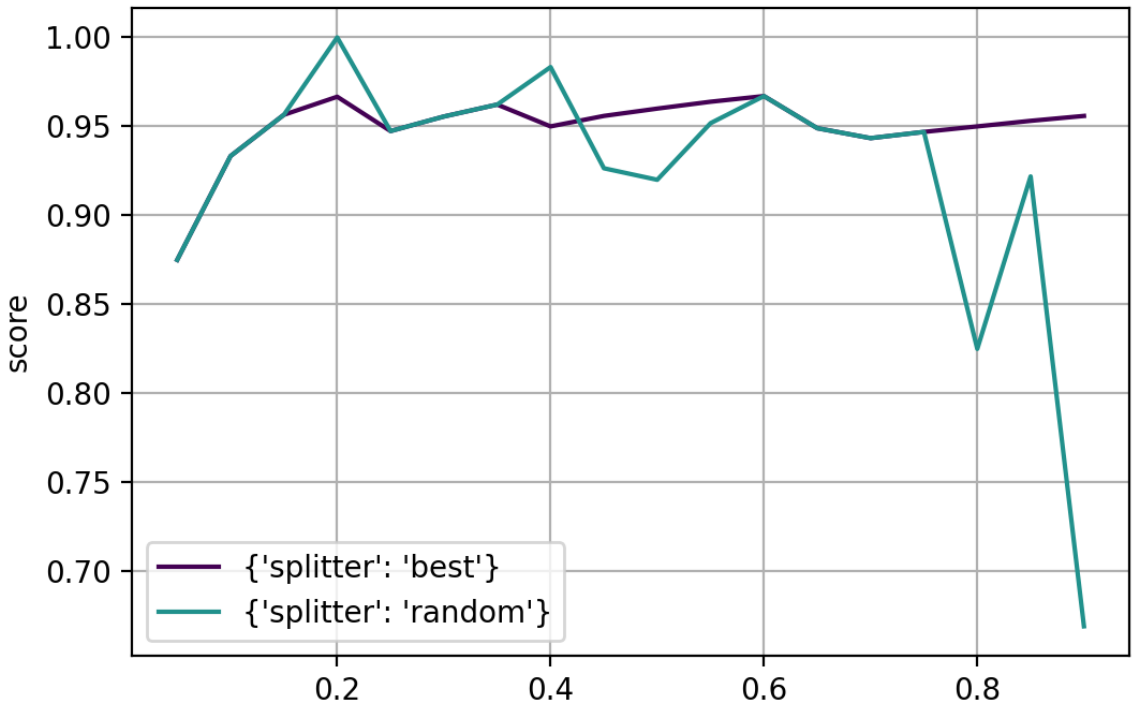

Рисунок 6 – splitter

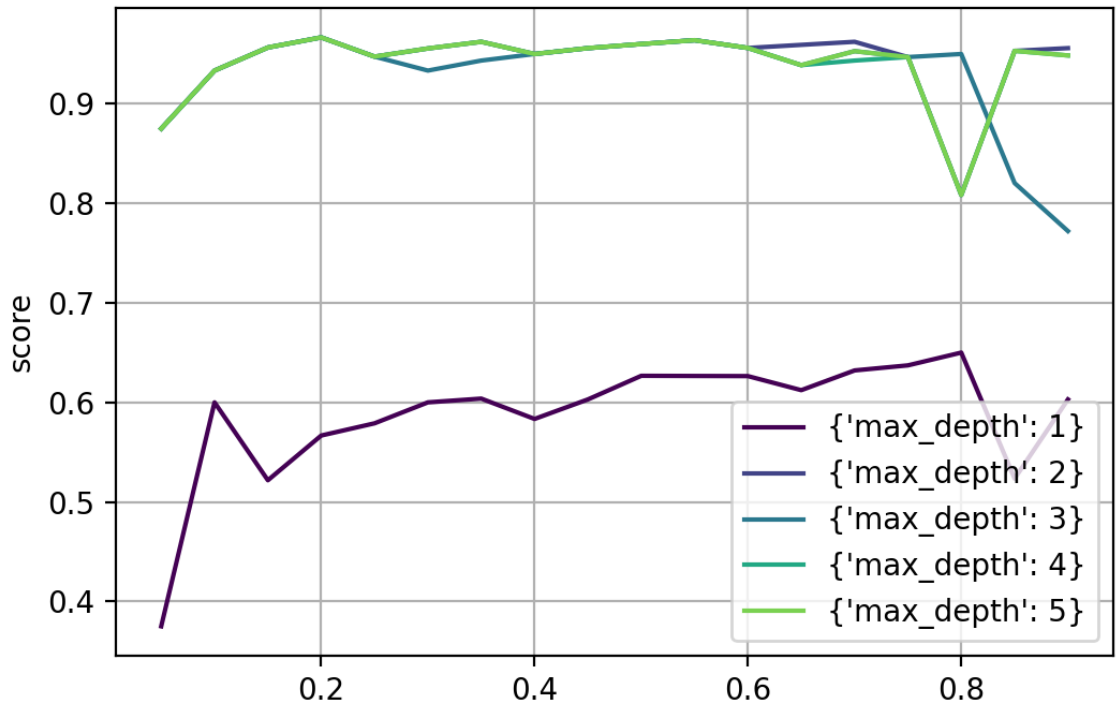

Рисунок 7 – max_depth

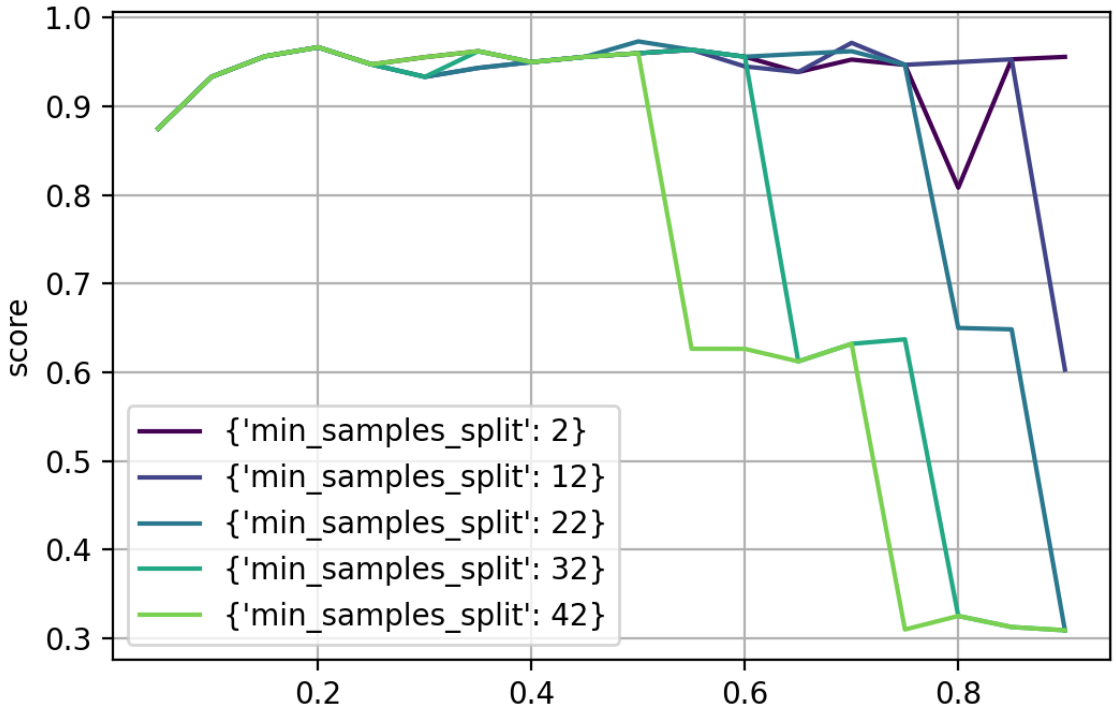

Рисунок 8 – min_samples_split

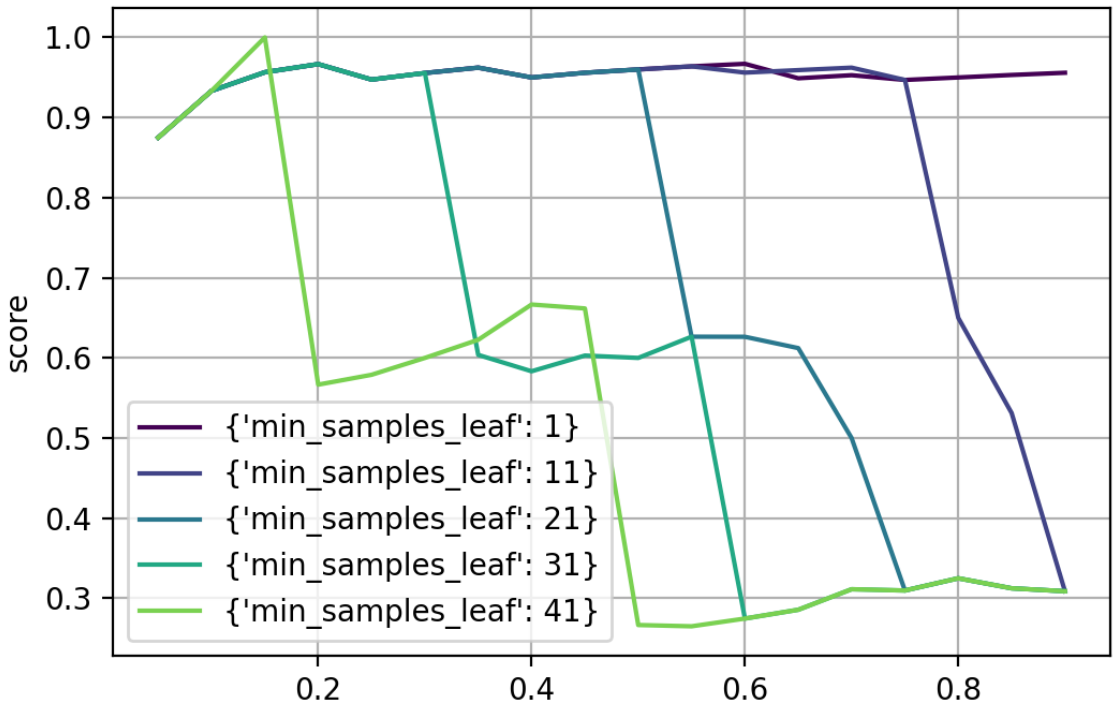

Рисунок 9 – min_samples_leaf

Как видно, энтропия и индекс Джини в данном случае показывают почти одинаковые результаты.

Случайная стратегия разделения справляется несколько хуже.

Сильно снижает качество классификации ограничение глубины дерева.

Увеличение `min_samples_split` и `min_samples_leaf` снижает точность классификации при ограничении набора.

Выводы

Произведено знакомство с наивным байесовским классификатором и деревьями решений в модуле Sklearn.

На данном наборе данных оба подхода хорошо работают даже при небольшом размере выборки для обучения.

Наивный байесовский классифкатор Бернулли плохо работает в случае не-бинарных данных.

Ограничение глубины дерева и слишком высокое число образов для разделения могут сильно снизить качество классификации деревом решений.