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

ОТЧЕТ

по лабораторной работе №7

по дисциплине «Машинное обучение»

Тема: Классификация (Байесовские методы, деревья)

Студент гр. 8303 _____ Гришин К. И.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами классификации из библиотеки *Sklearn*.

Ход выполнения работы

Загрузка данных

1. Скачать датасет: https://archive.ics.uci.edu/ml/datasets/iris.
2. Загрузить данные в датафрейм

	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

3. Данные отделены от меток
4. Метки преобразованы в числа
5. Выборка данных разбита на обучающую и тестовую

Байесовские методы

1. Проведена классификация наивным байесовским методом.

   Тестовая и обучающая выборки представляют собой исходные данные, поделенные пополам.

   Неправильно классифицировано 4 значения.

Атрибут	**Описание**
*class_count_*	Количество обучающих выборок, наблюдаемых в каждом классе
*class_prior_*	Вероятность каждого класса
*classes_*	Метки классов
*epsilon_*	Абсолютная аддитивная величина дисперсий
*sigma_*	Дисперсия каждого признака по классу
*theta_*	Среднее каждого признака по классу

2. Точность классификации *`score()`*.

$$`score(X_{test}, Y_{test})` = 0.99$$

я

3. График зависимости количества неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. *random_state = 830303*. (рис. 1)

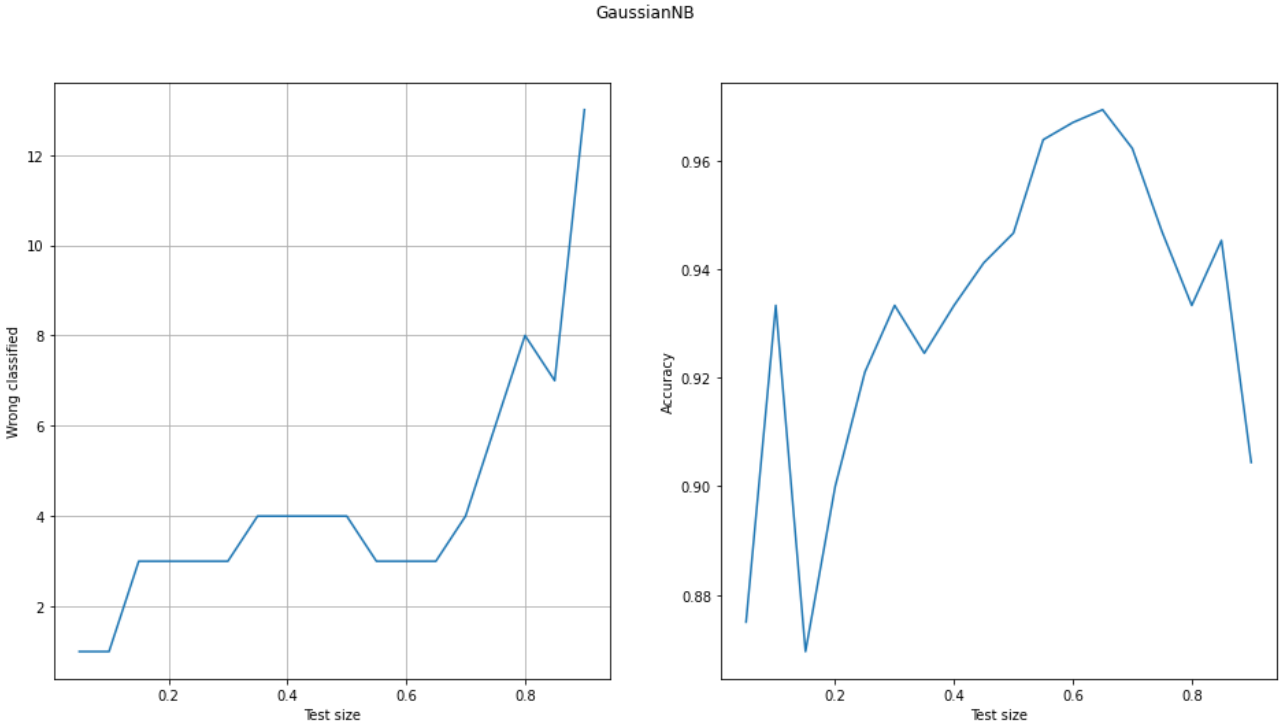

Рисунок 1. Зависимость количества неправильно классифицированных данных и точности классификации от размера тестовой выборки *GaussianNB*.

Точность классификации остается выше 90% на большей части выборки. При размере тестовой выборке от 0.2, точность классификатора стабильно увеличивается.

Это связано с нормальным распределением данных в классах, на основе которых ведется расчет в GaussianNB.

4. Проведена классификация с использованием *MultinomialNB*, *ComplementNB*, *BernoulliNB* (рис. 2, 3, 4 соотв.).

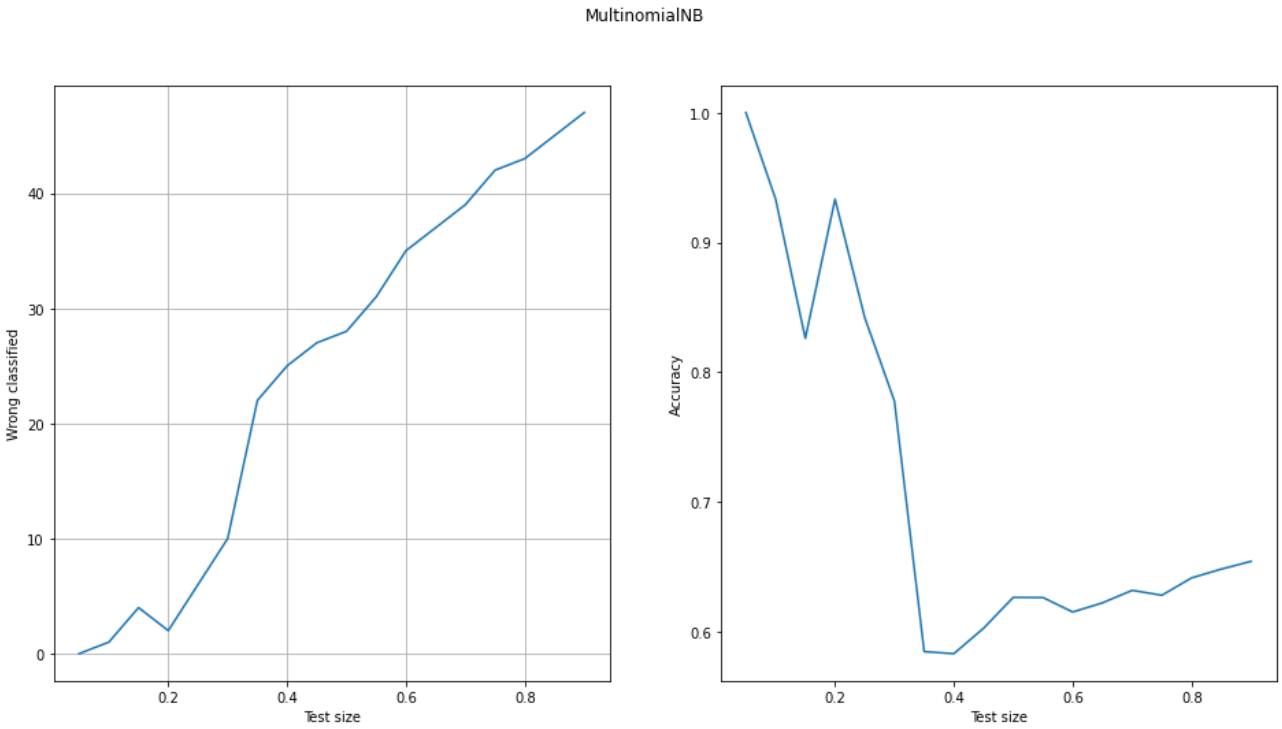

Рисунок 2. Зависимость количества неправильно классифицированных данных и точности классификации от размера тестовой выборки *MultinomialNB*.

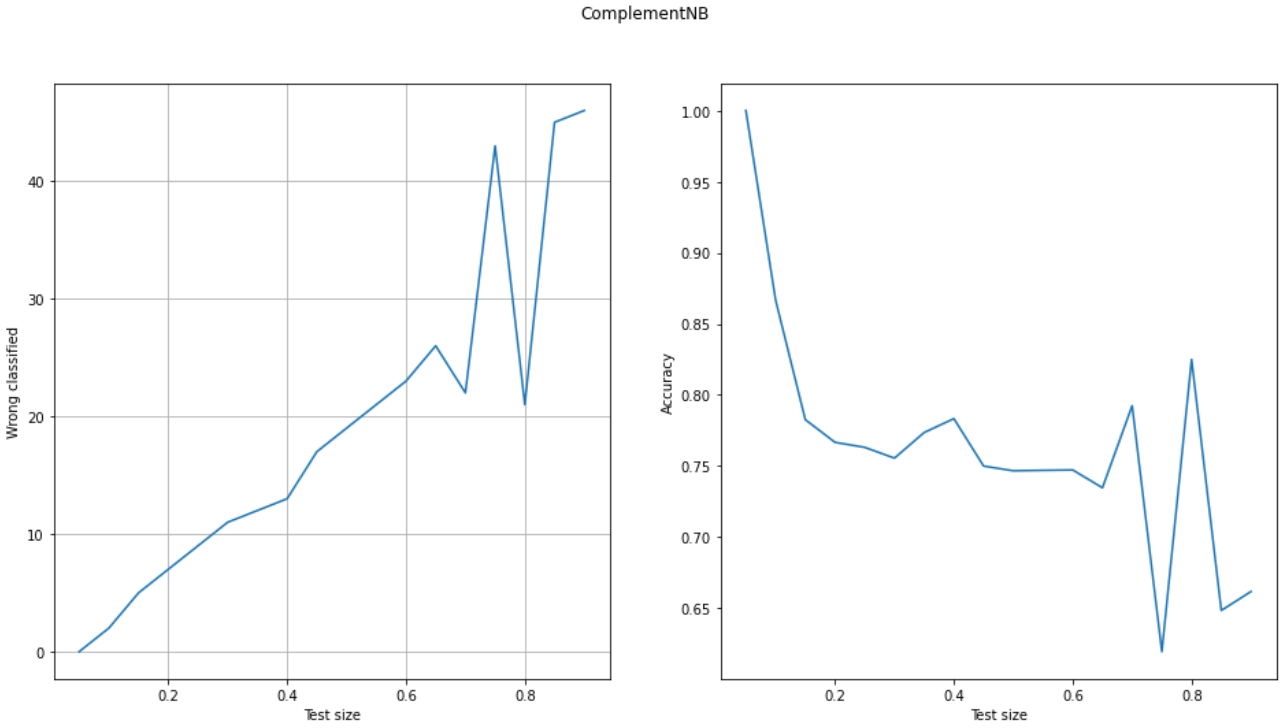

Рисунок 3. Зависимость количества неправильно классифицированных данных и точности классификации от размера тестовой выборки *ComplementNB*.

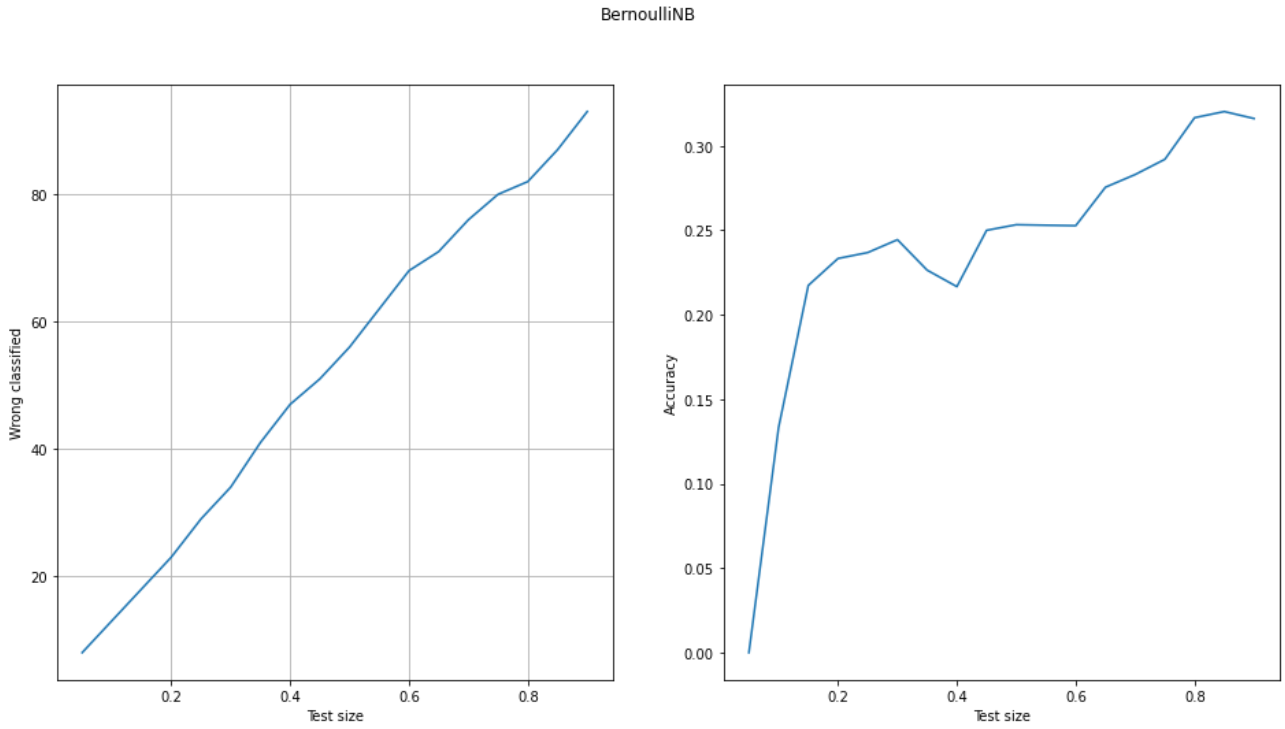

Рисунок 4. Зависимость количества неправильно классифицированных данных и точности классификации от размера тестовой выборки *BernoulliNB*.

MultinomialNB – полиномиальное функция появления наблюдения в классе.

ComplementNB – важное отличие в способе определения вероятности принадлежности классу. В отличие от обычного NB, здесь ведется поиск минимума в принадлежности другим классам.

BernoulliNB – реализация NB, где данные представлены многомерными векторами Бернулли (содержат двоичные данные).

Лучший результат показал *GaussianNB*, поскольку данные действительно нормально распределены. *MultinomialNB* и *ComplementNB*, показали худший результат, поскольку оперируют полиномиальным распределением.

Худший результат показал *BernoulliNB*, т.к. он оперирует данными отличными от предоставленных.

Классифицирующие деревья

1. Проведена классификация при помощи деревьев на тех е данных. Неправильно классифицировано 4 значения

2. Точность классификации `*score()*`.

$$`score( )` = 0.95$$

3. Характеристики деревьев.

$$get\_n\_leaves() = 5; \quad get\_depth() = 4$$

4. Полученное дерево (рис. 5).

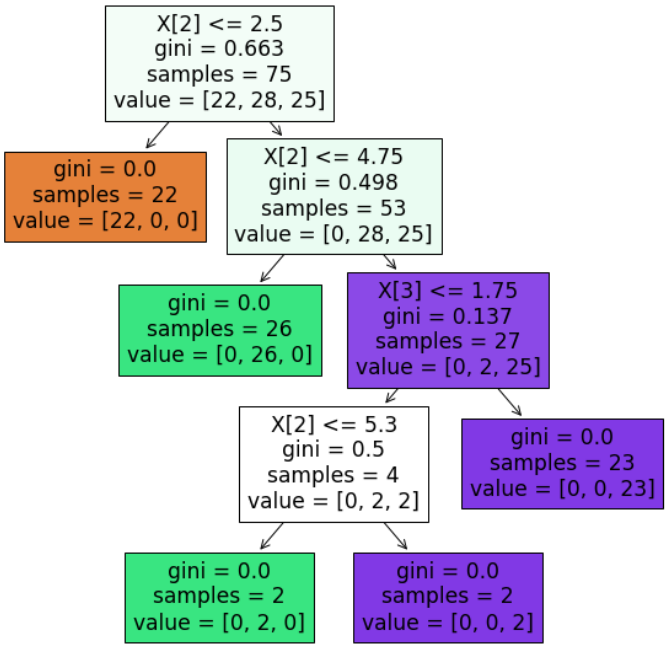

Рисунок 5. Дерево *DecisionTreeClassifier*.

Для каждого узла указываются: условие разбиения по признаку, значение загрязненности, количество наблюдений, распределение наблюдений по классам. В листах условие разбиения отсутствует.

5. График зависимости количества неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. *random_state = 830303* (рис. 6).

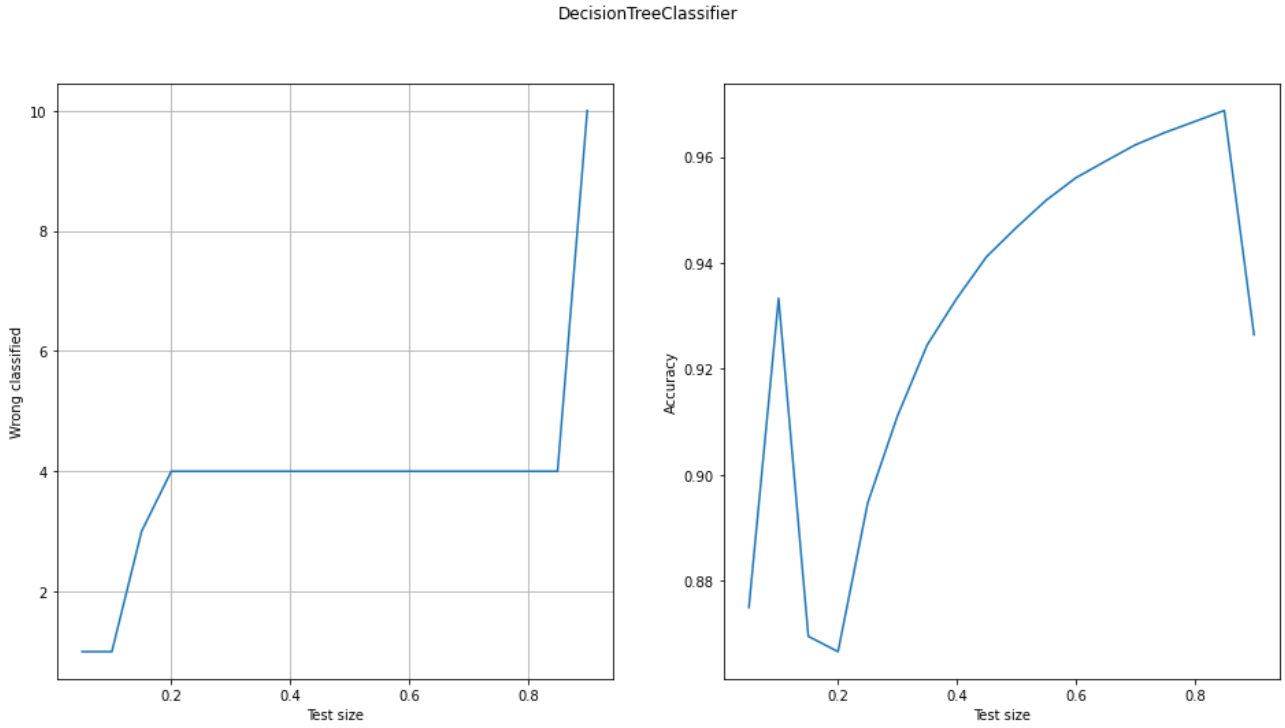

Рисунок 6. Зависимость количества неправильно классифицированных данных и точности классификации от размера тестовой выборки *DecisionTreeClassifier*.

Наблюдается слабая зависимость между качеством классификации и размером выборки, что говорит, о хорошей классифицируемости данных выборки.

6. Работа классифицирующего дерева при различных параметрах *criterion*, *splitter*, *max_depth*, *min_samples_split*, *min_samples_leaf*.

criterion (рис. 7)

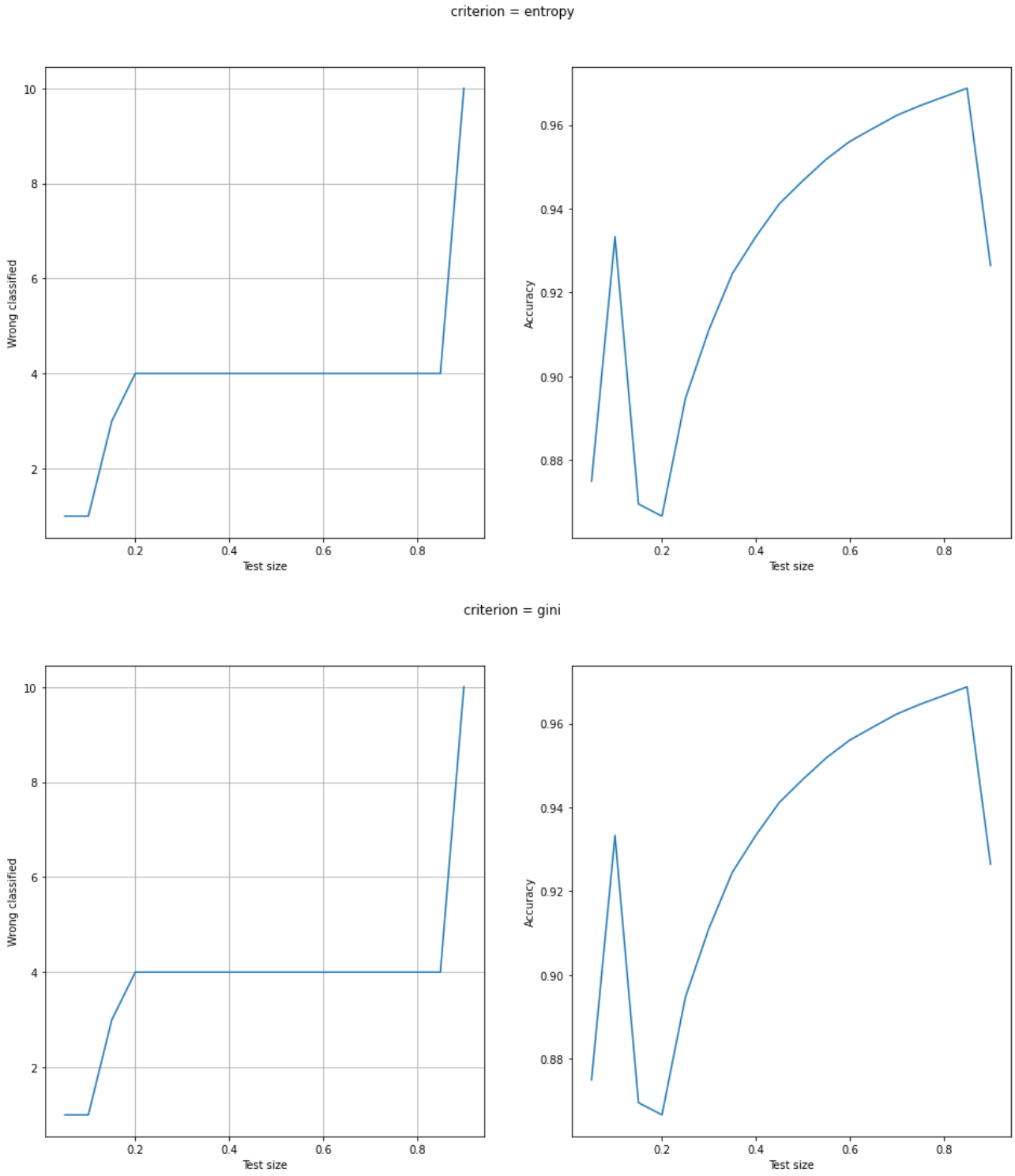

Рисунок 7. *DecissionTreeClassifier(criterion=)*

splitter (рис. 8)

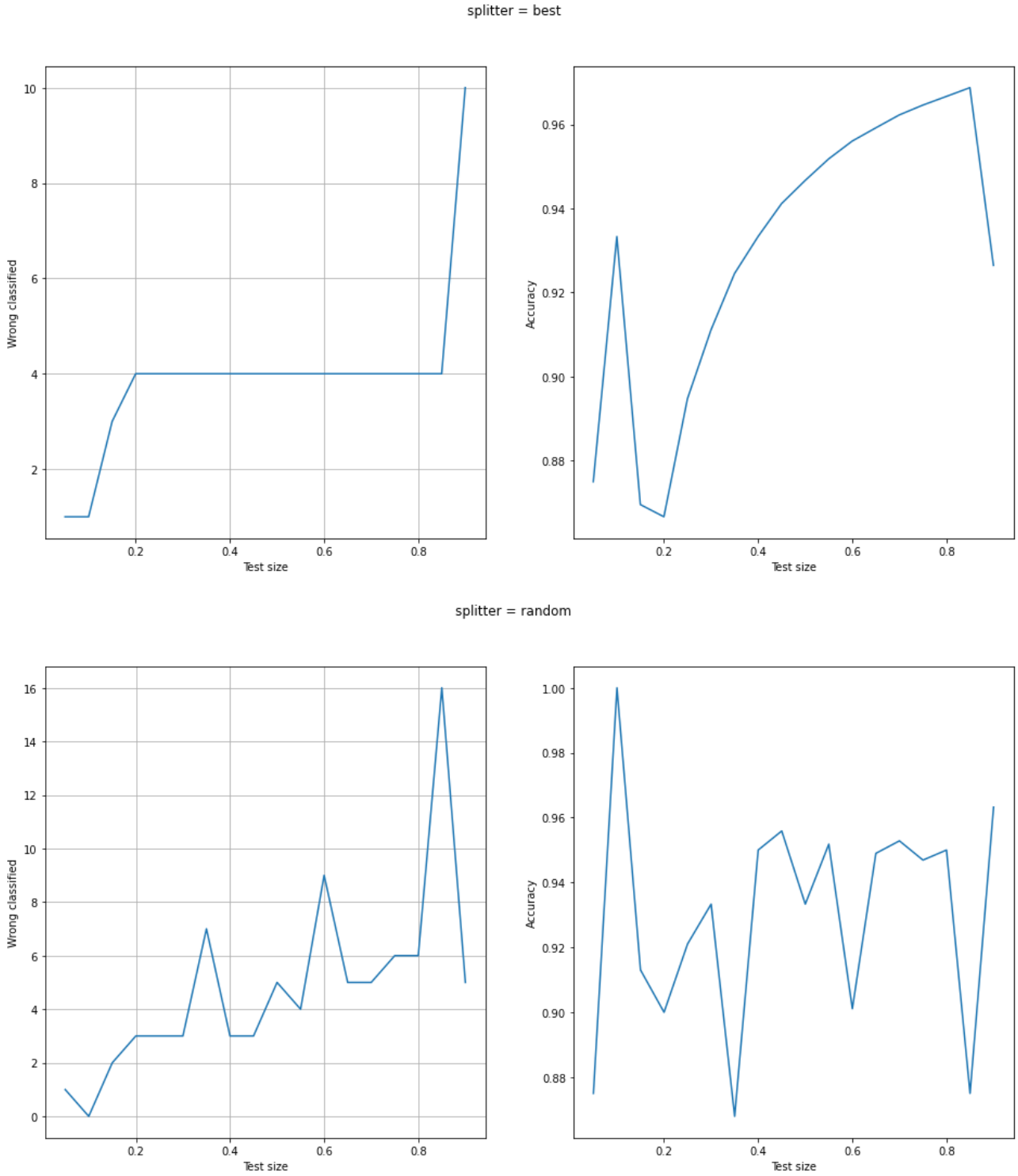

Рисунок 8. *DecissionTreeClassifier(splitter=)*.

max_depth (рис. 9, 10)

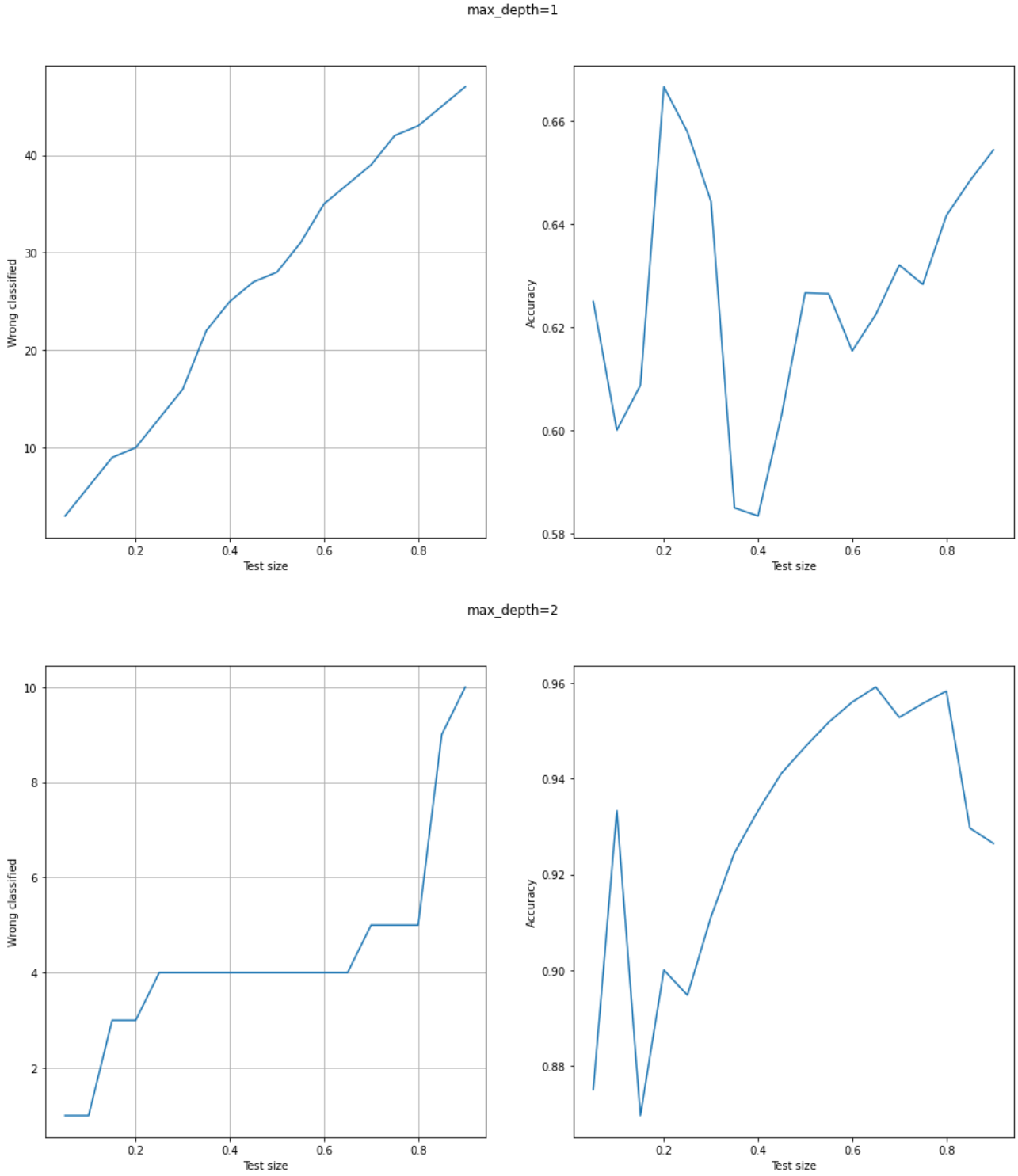

Рисунок 9. *DecissionTreeClassifier(max_depth=).*

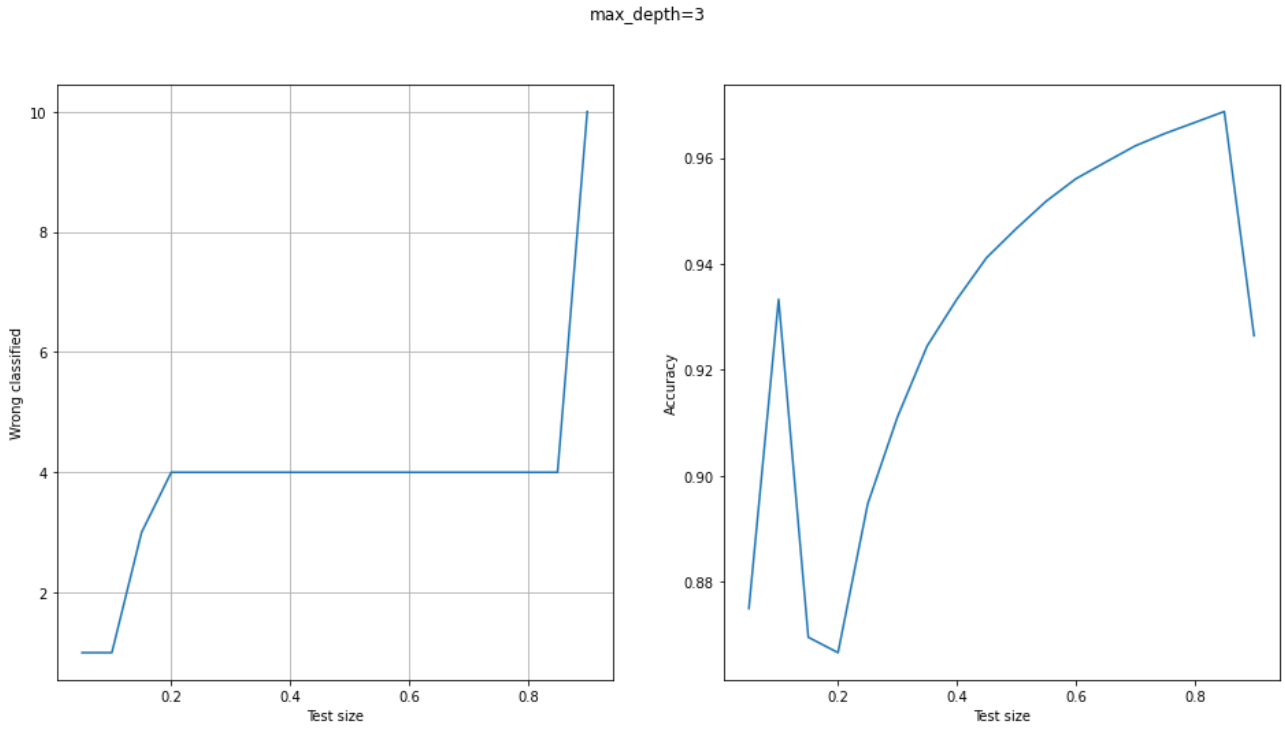

Рисунок 10. *DecissionTreeClassifier(max_depth=)*.

min_samples_split (11, 12, 13)

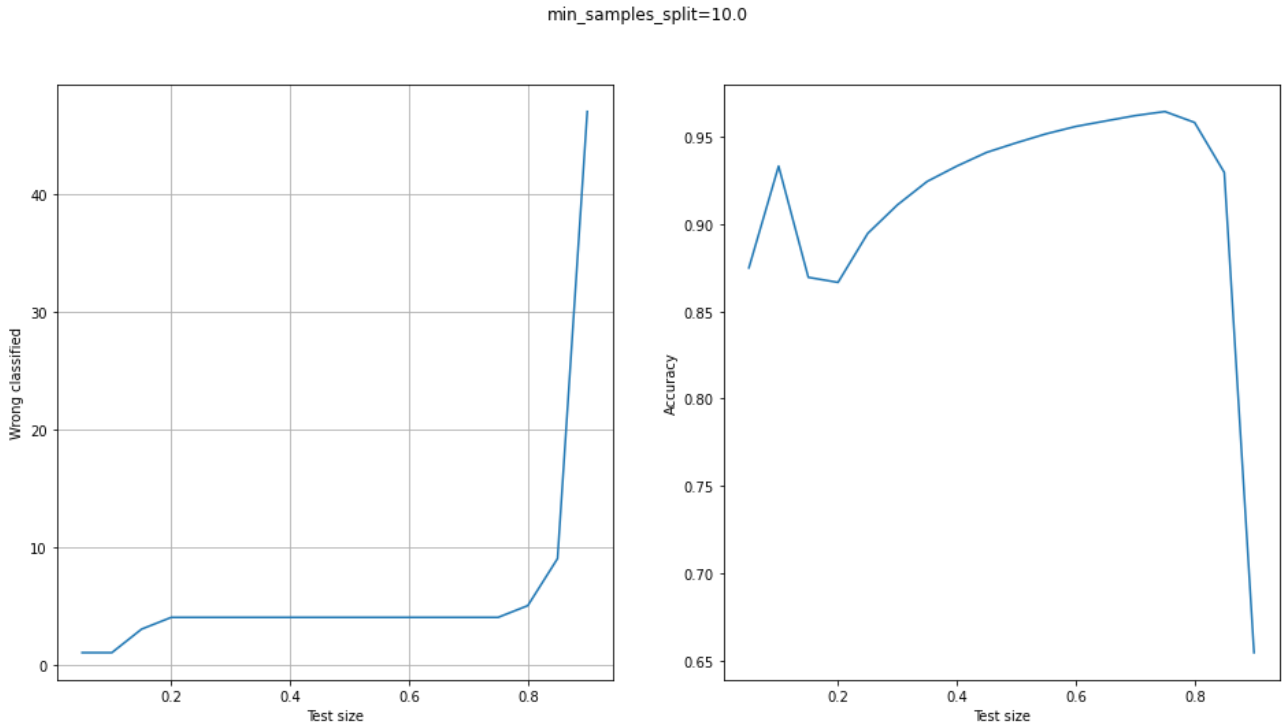

Рисунок 11. *DecissionTreeClassifier(min_samples_split=)*.

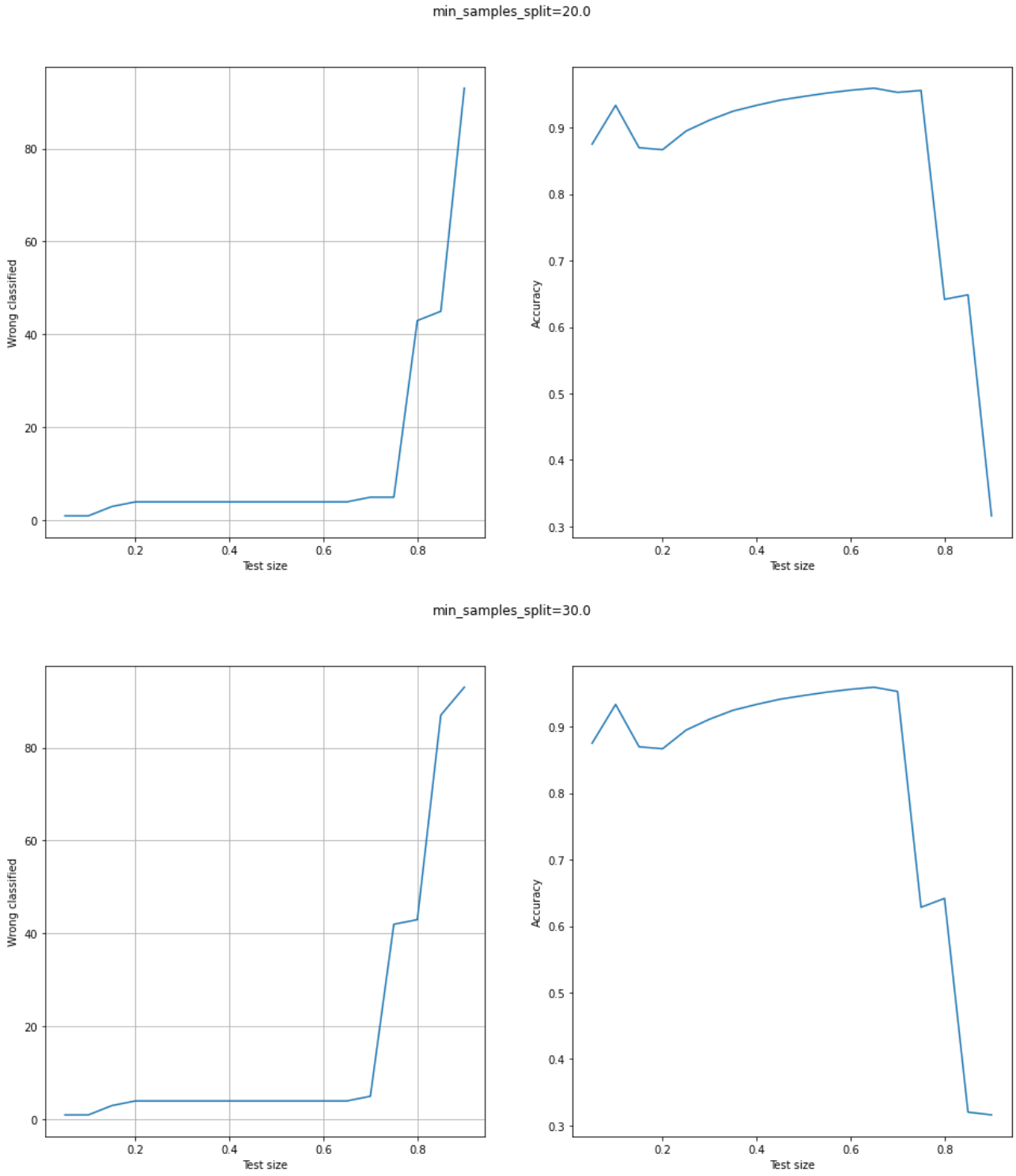

Рисунок 12. *DecissionTreeClassifier(min_samples_split=).*

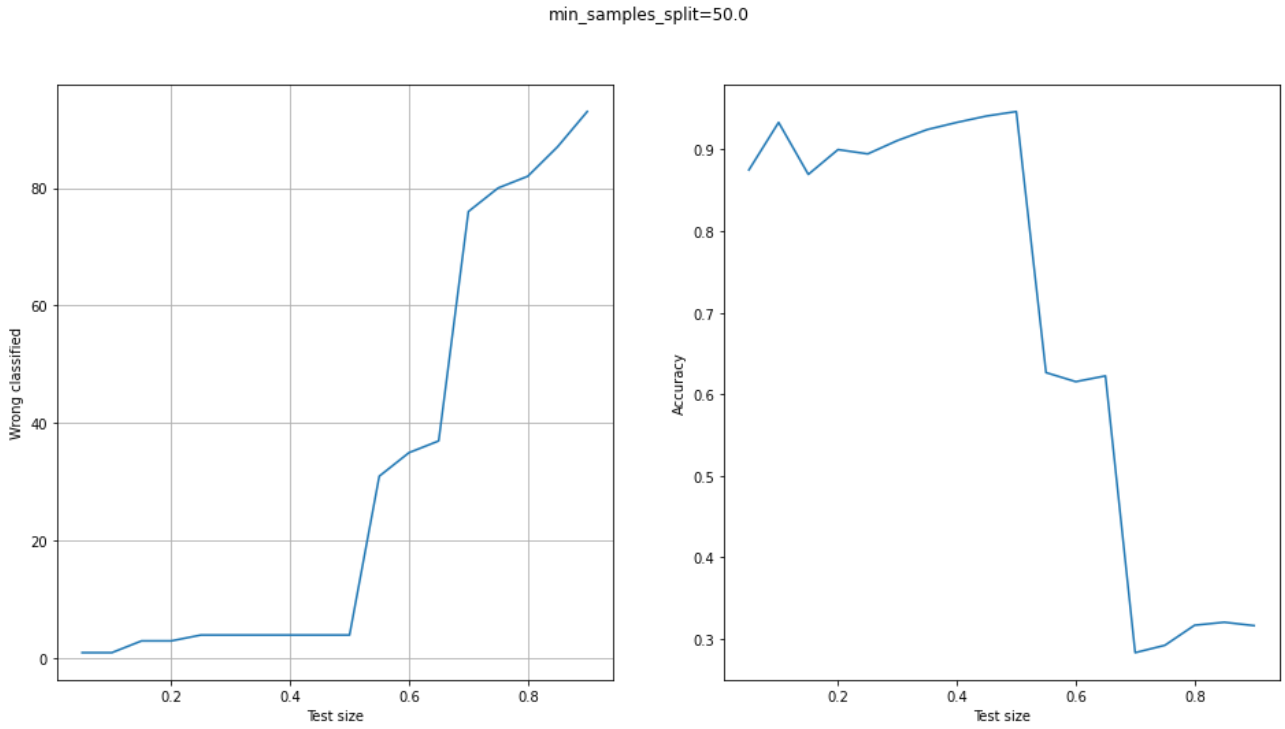

Рисунок 13. *DecissionTreeClassifier(min_samples_split=).*

min_samples_leaf (рис 14, 15)

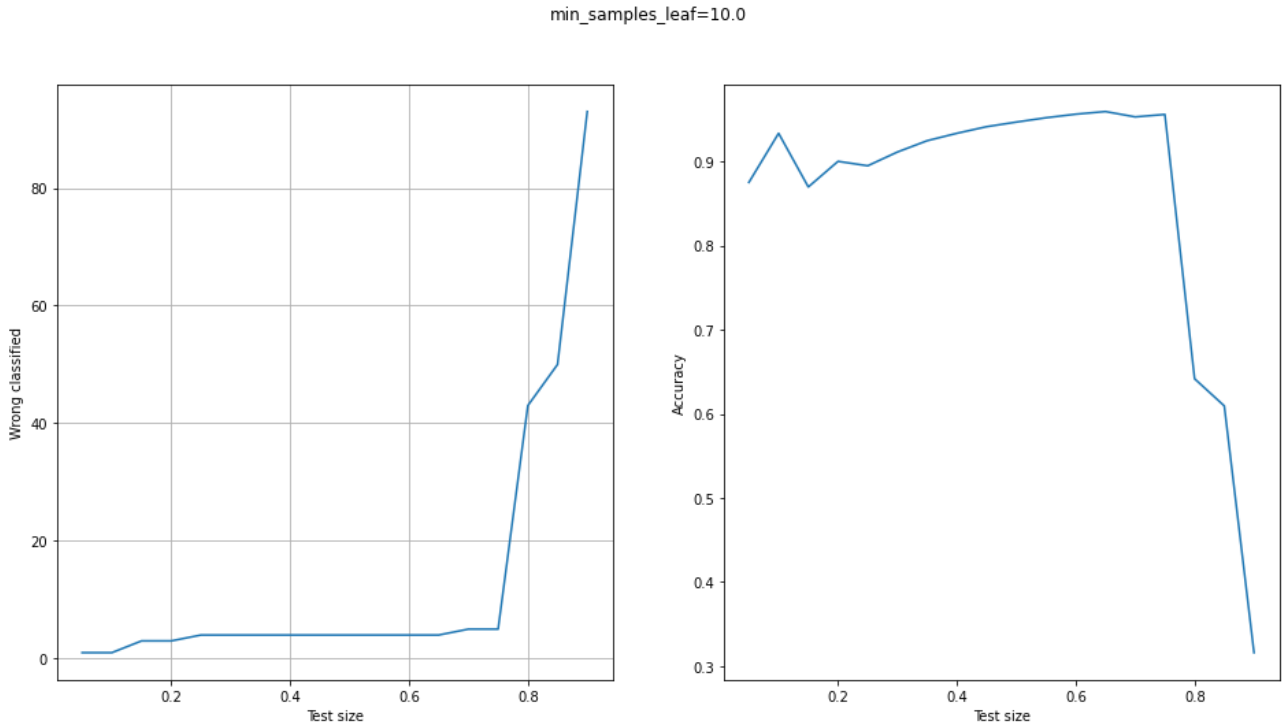

Рисунок 14. *DecissionTreeClassifier(min_samples_leaf=).*

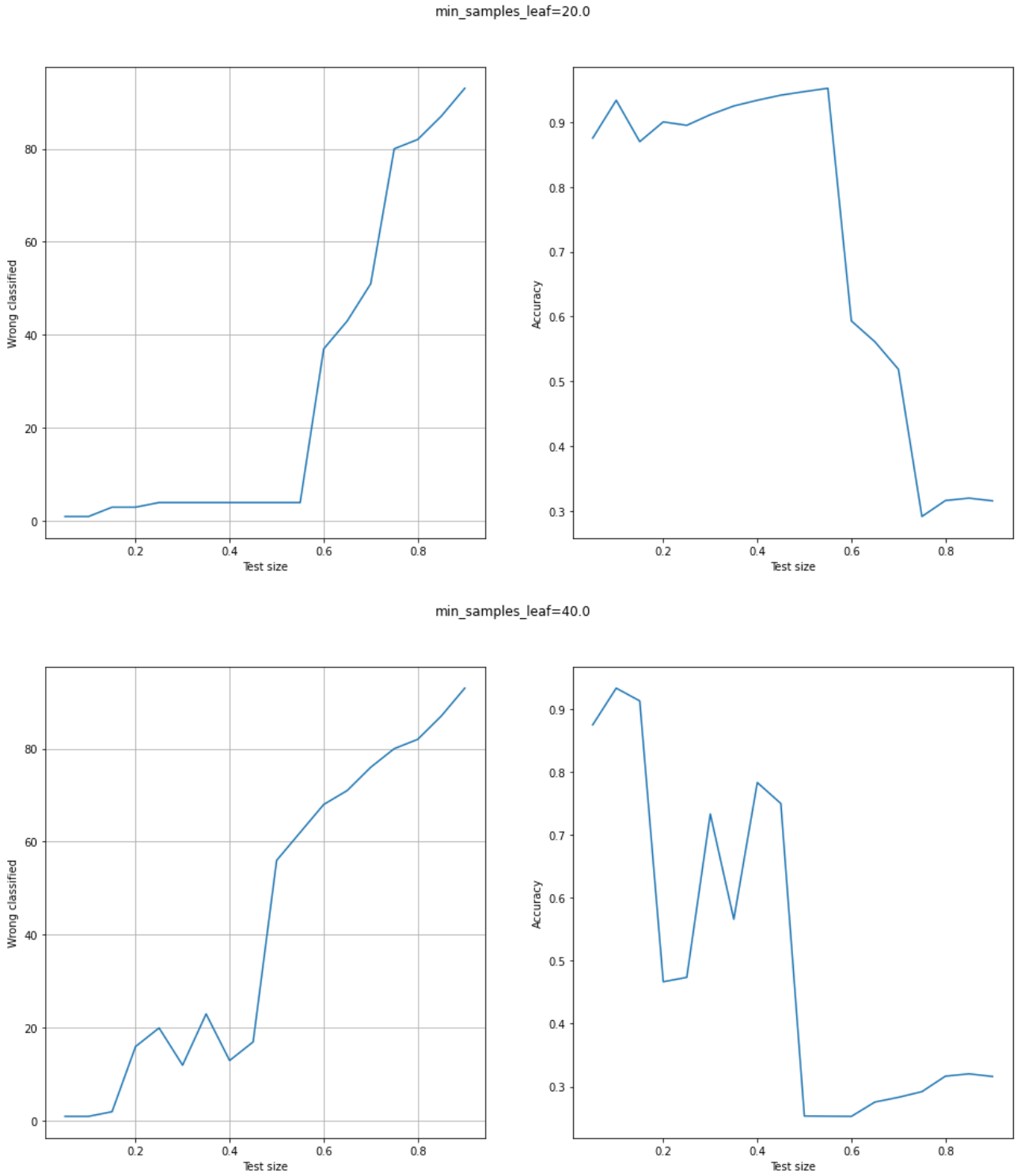

Рисунок 15. *DecissionTreeClassifier(min_samples_leaf=).*

Параметр	Описание
criterion	Критерий определения загрязненности узла. Не повлиял на результаты.
splitter	Стратегия разбиения узла. Может быть лучшим или случайным. Случайное переобучает дерево и ведет к ухудшению классификации.
max_depth	Максимальная глубина дерева. При уменьшении ведет к ухудшению результата.
min_sampes_split	Минимальное количество данных узла. Если данных меньше указанных, то узел не может быть разделен. При сильном увеличении сильно страдает точность
min_samples_leaf	Минимальное количество данных наследника. Не может быть создан наследник с количеством данных, меньше указанного. При сильном увеличении сильно страдает точность

Вывод

В ходе лабораторной работы исследованы методы классификации: *NaiveBayes* и *DecisionTreeClassifier*.

NaiveBayes – наивный метод классификации, опирающийся на независимость признаков. Основной задачей является поиск вероятности попадания в класс $p(f_1, \dots, f_n|c) = \prod_{i=1}^{n} p(f_i|c)$. В обычном NB выбирается класс, который имеет наибольшую вероятность. Однако существуют различные модификации. По различному может определятся выбор $p(f_i|c)$ и выбор наиболее подходящего класса.

DecisionTreeClassifier – метод классификации данных, основанный на по строении дерева вариантов. Используя критерий загрязненности, выбираются наиболее подходящие места разделения узлов (изначально все данные представляют собой один большой узел). В какой-то момент, загрязненность становится слишком мала, или дальнейшее разделение нерационально, полученный узел называется определяющим класс.