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. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

Загрузка данных

Датасет загружен в датафрейм. Выделены данные и их метки, тексты меток преобразованы к числам. Выборка разбита на обучающую и тестовую train_test_split.

Байесовские методы

Проведена классификация наблюдений наивным байесовским методом. Выявлено 4 неправильно классифицированных наблюдения. Атрибуты классификатора представлены в табл. 1.

Таблица 1 – Атрибуты GaussianNB

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

Точность классификации составляет 96%.

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

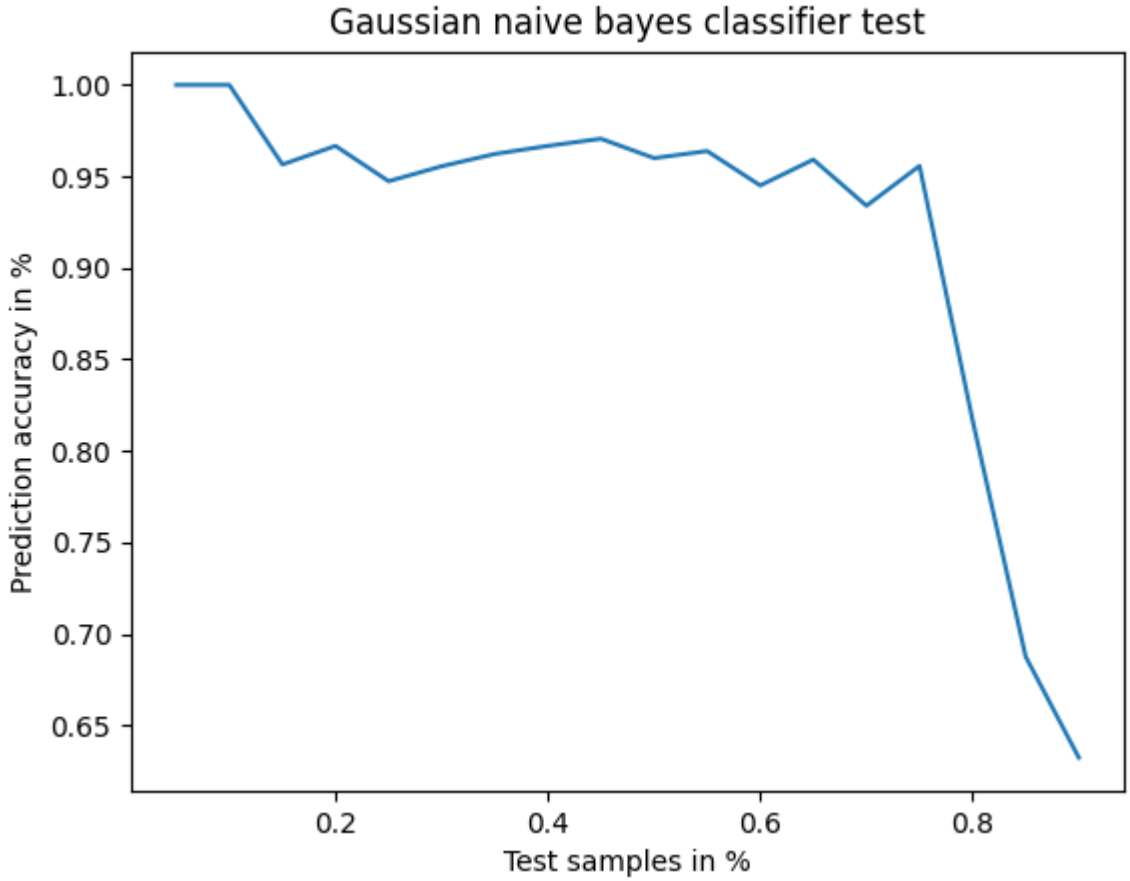

Рисунок 1 – Классификация GaussianNB

Классификация проведена с помощью MultinominalNB, ComplementNB, BernoulluNB. Результат представлен на рис. 2 – 4.

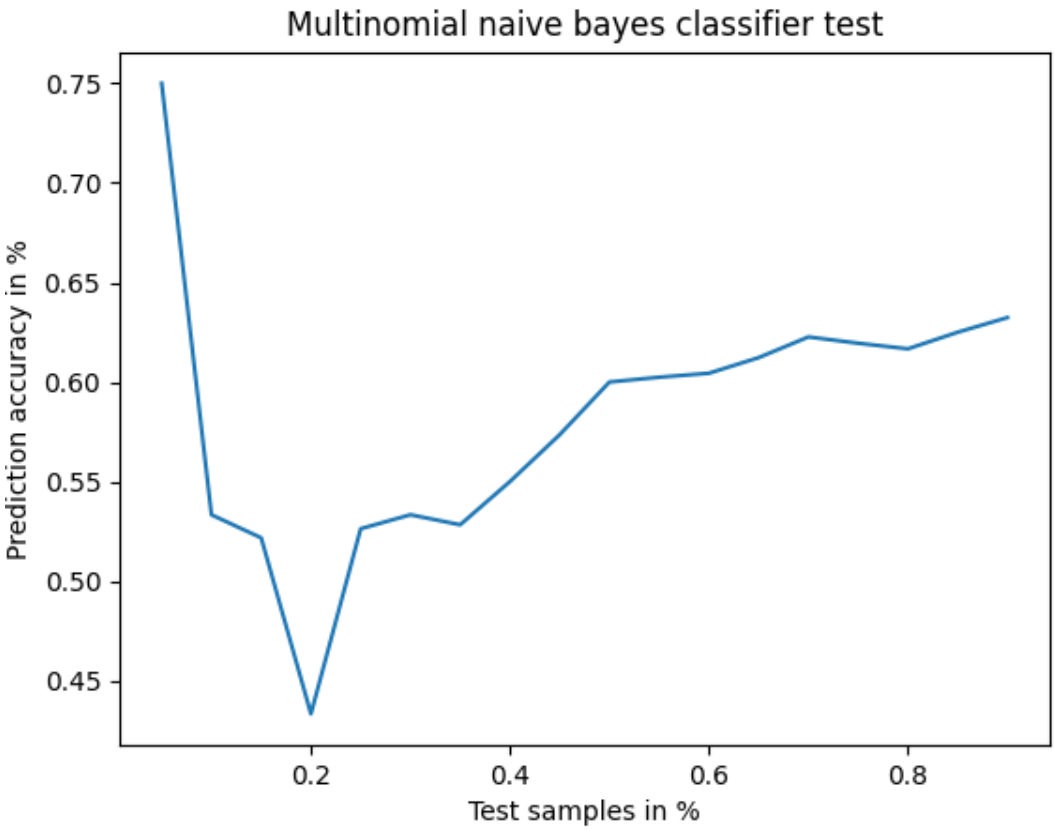

Рисунок 2 – Классификация MultinomialNB

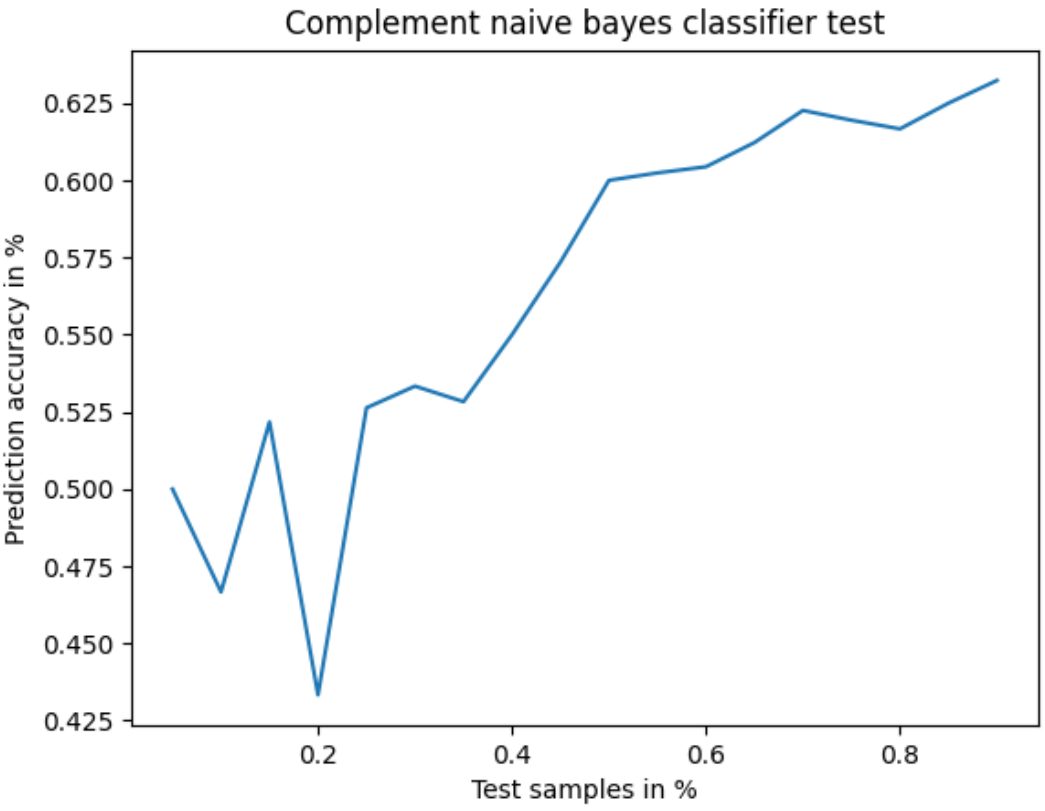

Рисунок 3 – Классификация ComplementNB

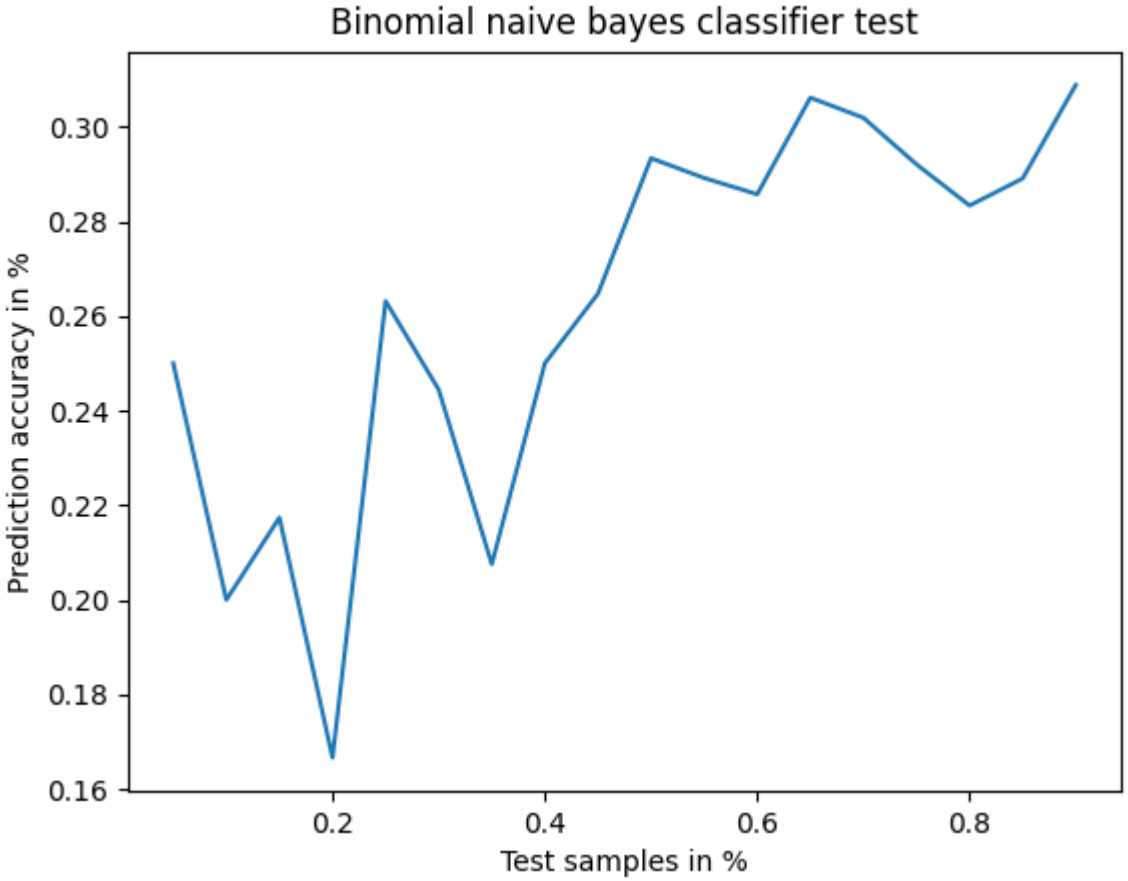

Рисунок 4 – Классификация BinomialNB

MultinominalNB – полиномиальный наивный байесовский классификатор, подходит для классификации с дискретными признаками (например, подсчет слов для классификации текста). MultinominalNB реализует наивный алгоритм Байеса для полиномиально распределенных данных. Распределение для каждого класса параметризируется векторами, содержащими вероятности вхождения признаков в элемент выборки, соответствующий данному классу.

ComplementNB – адаптация MultinominalNB, подходит для несбалансированных наборов данных. В частности, CNB использует статистику из дополнения каждого класса для вычисления весов модели. ComplementNB часто превосходит MultinominalNB в задачах классификации текста.

BernoulluNB – как и MultinominalNB, этот классификатор подходит для дискретных данных. Разница в том, что в то время, как MultinominalNB работает с подсчетом вхождений, BernoulluNB предназначен для двоичных/логических признаков.

Классифицирующие деревья

Проведена классификация наблюдений с помощью деревьев решений на тех же данных. Выявлено 5 неправильно классифицированных наблюдения.

Точность классификации составляет 93%.

Получившееся дерево имеет глубину, равную 5, и 7 листа.

Дерево продемонстрировано на рис. 5.

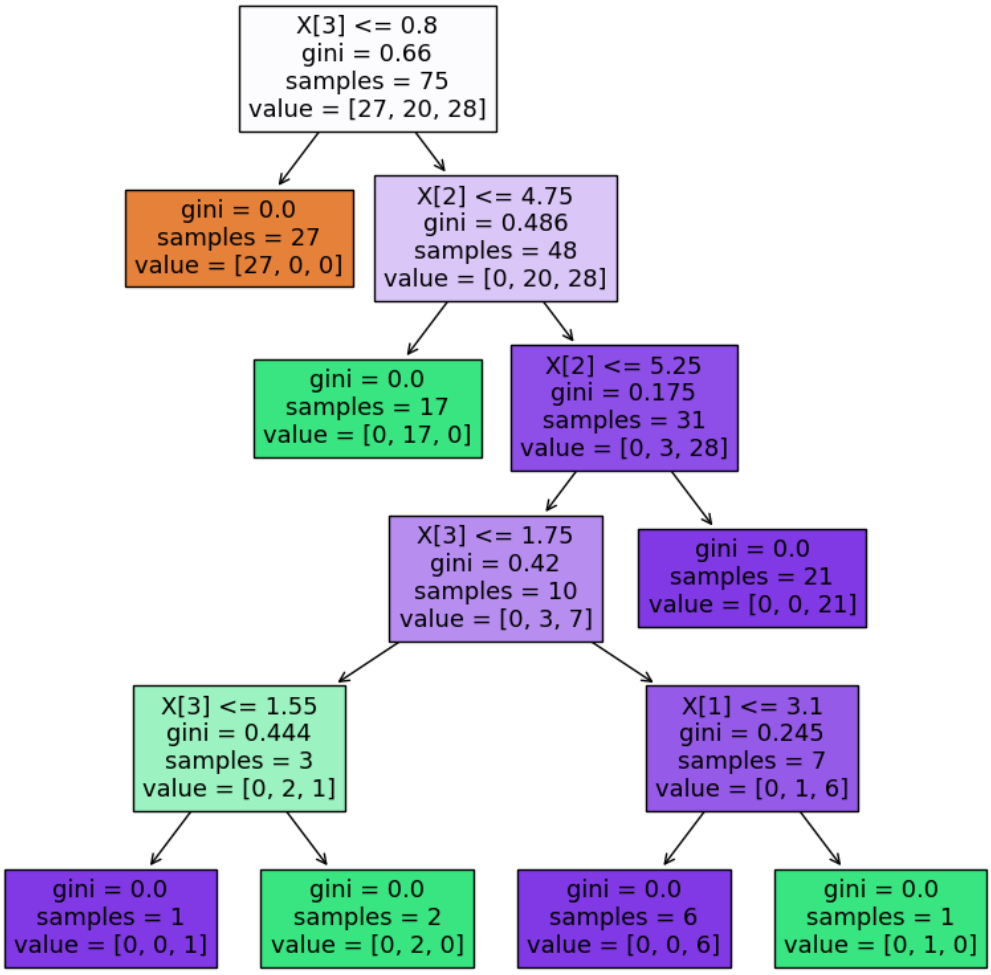

Рисунок 5 – Дерево решений для классификации

Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. График представлен на рис. 6.

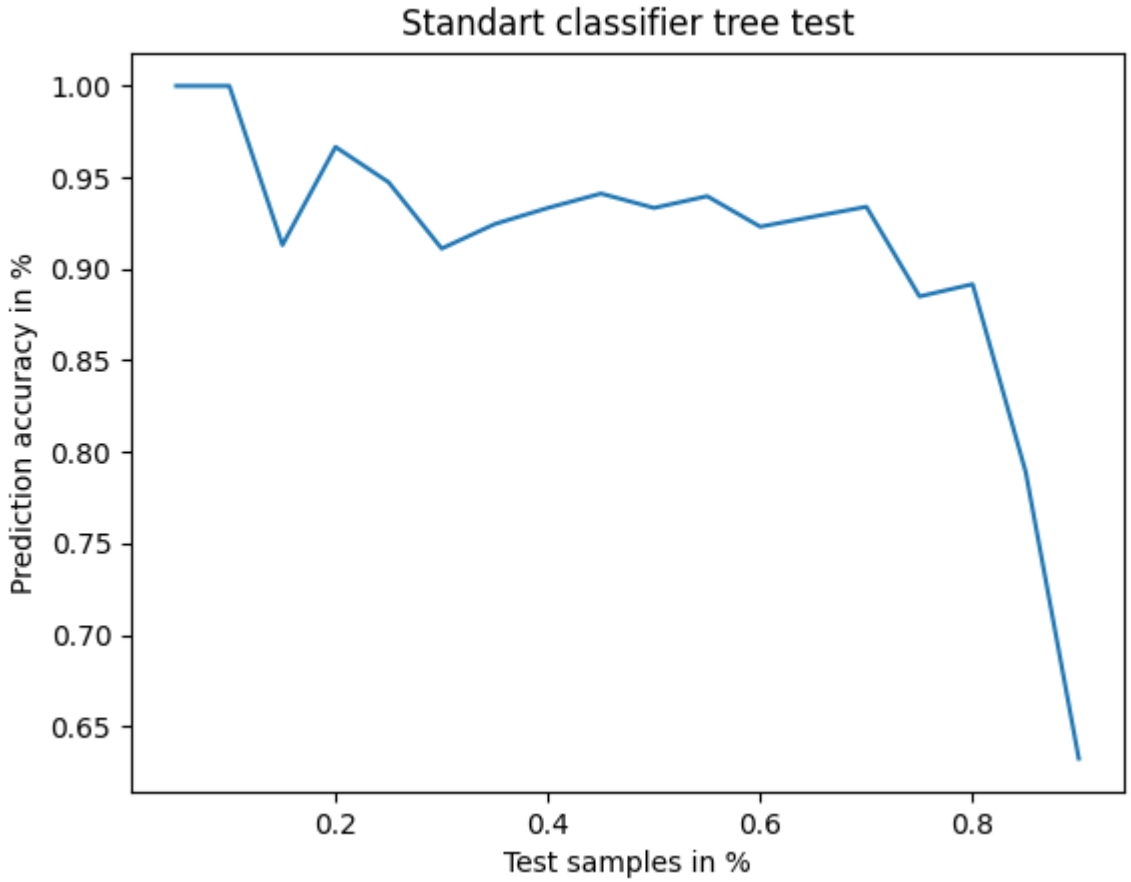

Рисунок 6 – Классификация DecisionTreeClassifier

Исследованы параметры DecisionTreeClassifier, результаты представлены в табл. 2. и на рис. 7 – 11.

Таблица 2 – Парметры DecisionTreeClassifier

Параметр	Описание
criterion	Функция измерения качества разбиения. Поддерживается индекс Джини и энтропия.
splitter	Стратегия, используемая для выбора разбиения на каждом узле. Поддерживается выбор наилучшего разбиения и случайный выбор.

max_depth	Максимальная глубина дерева. Если None, то узлы расширяются до тех пор, пока все листья не станут чистыми или пока все листья не будут содержать менее min_samples_split выборок.
min_samples_split	Минимальное количество выборок, необходимых для разделения внутреннего узла.
min_samples_leaf	Минимальное количество выборок, которое требуется для конечного узла. Точка разделения на любой глубине будет учитываться только в том случае, если она оставляет не менее min_samples_leaf обучающих выборок в каждой из левой и правой ветвей.

Рисунок 7 – Тест параметра criterion

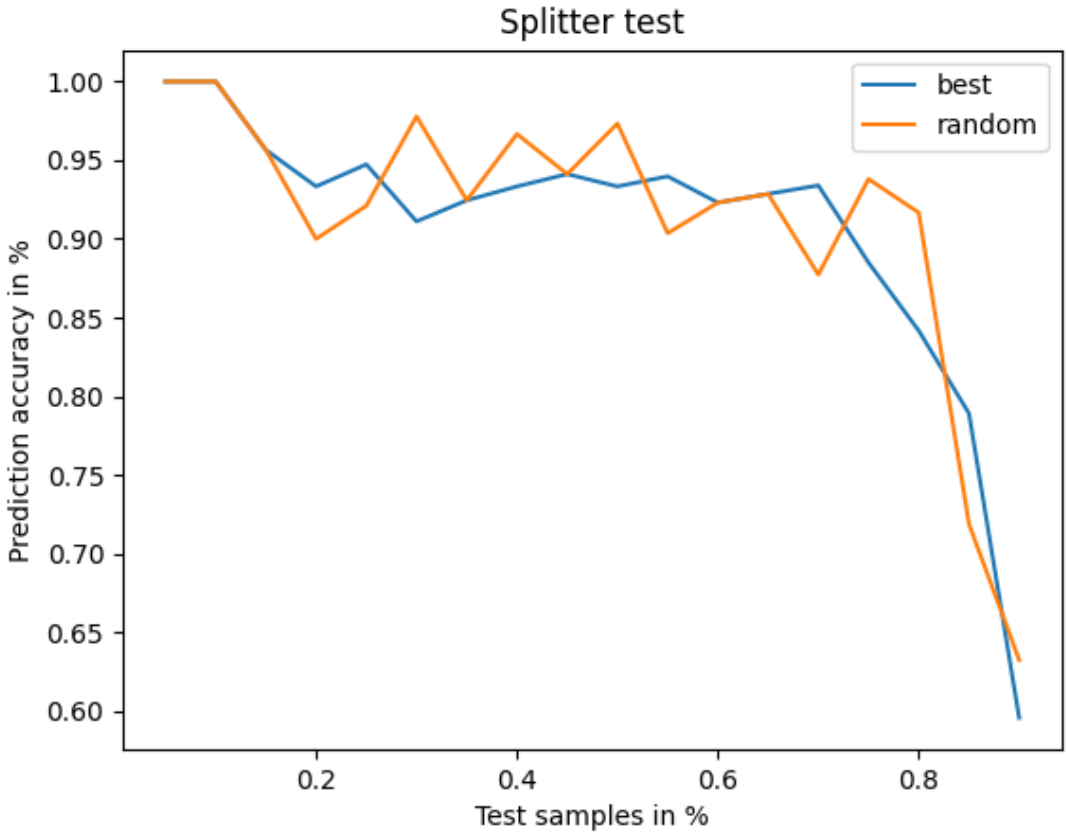

Рисунок 8 – Тест параметра splitter

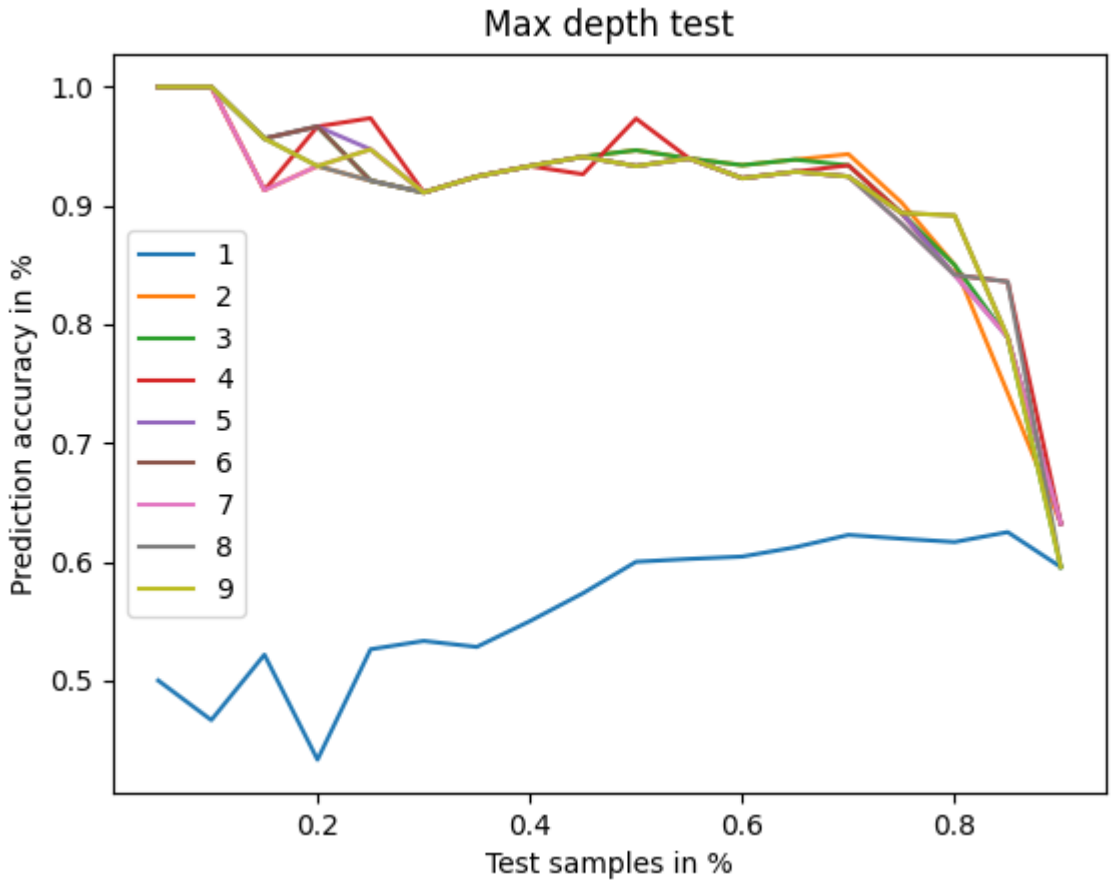

Рисунок 9 – Тест параметра max_depth

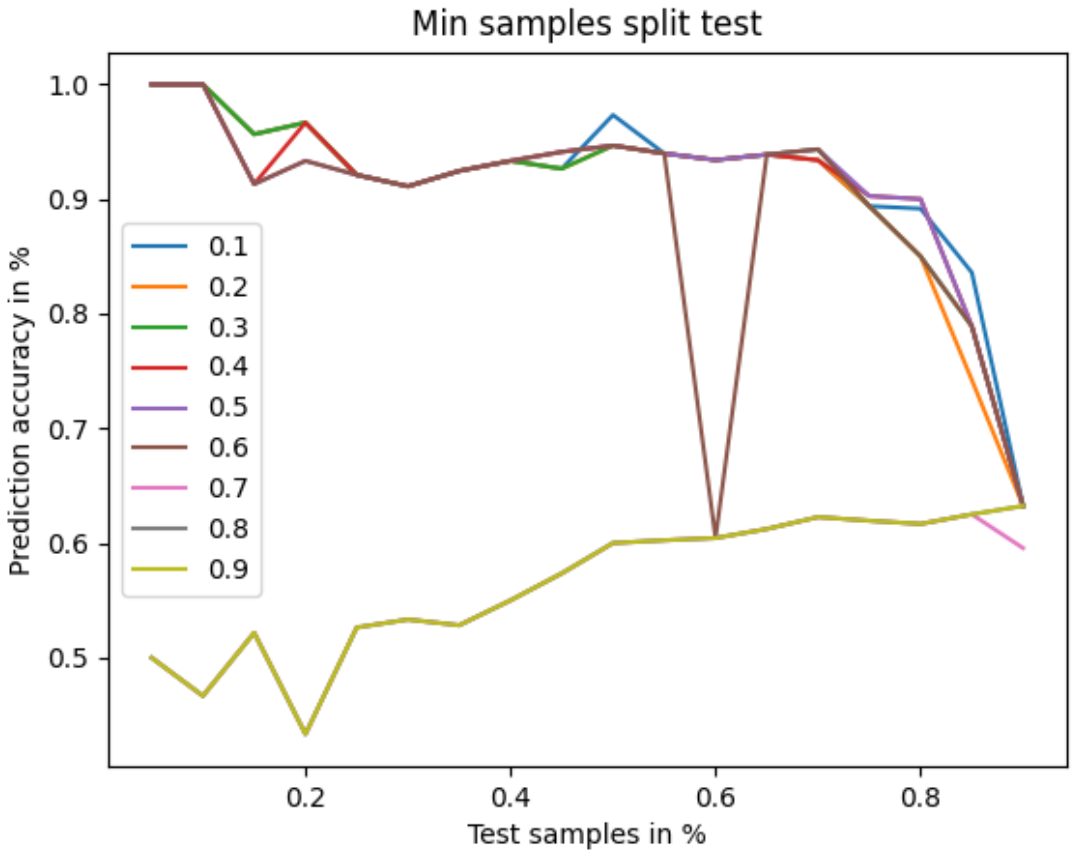

Рисунок 10 – Тест параметра min_samples_split

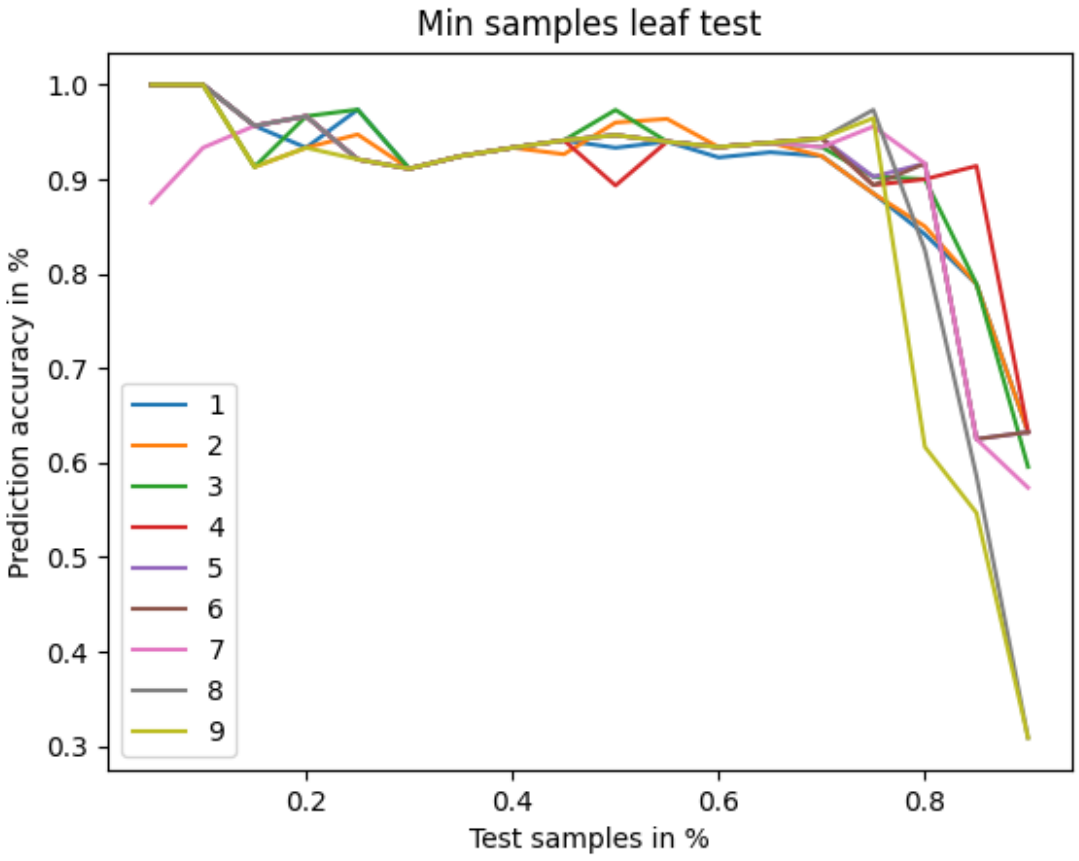

Рисунок 11 – Тест параметра min_samples_split

Выводы

В ходе лабораторной работы рассмотрены такие методы классификации модуля Sklearn, как GaussianNB, MultinominalNB, ComplementNB, BernoulluNB и DecisionTreeClassifier.