**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**КАФЕДРА САПР**

**ОТЧЕТ**

**по Лабораторной работе № 3**

**«ДЕРЕВЬЯ И ЛЕСА РЕШЕНИЙ»**

**по дисциплине «Автоматизация схематического проектирования»**

|  |  |  |
| --- | --- | --- |
| Студенты гр. 0302 |  | Сухарев Л. А.. |
| Преподаватель |  | Боброва Ю. О. |

Санкт-Петербург

2024

**Цель:** реализация классификатора на основе дерева принятия решений и исследование его свойств.

1. **ОСНОВЫНЕ ТЕОРЕТИЧЕСКИЕ ПОЛОЖЕНИЯ**

Дерево решений (распознающее дерево, recognition или decision tree) –распознаватели вида, при котором для распознаваемого объекта проводится конечная последовательность сравнений значений его признаков с константами на равенство или неравенство, причем от результатов каждого сравнения зависят дальнейшие действия – продолжать сравнивать с чем-то еще или давать ответ распознавания. То есть распознавание можно представить как двоичное дерево вложенных операторов вида:

if x[j] ?? d[k] then …

else…

завершающееся листьями:

return res[h]

где x-массив признаков, d-массив пороговых значений, res-массив возможных ответов, знаком ?? обозначена операция сравнения.

Деревья строятся при помощи обучения с учителем. В качестве обучающего набора данных используется множество наблюдений, для которых предварительно задана метка класса.

Структурно дерево решений состоит из объектов двух типов — узлов (node) и листьев (leaf). В узлах расположены решающие правила и подмножества наблюдений, которые им удовлетворяют. В листьях содержатся классифицированные деревом наблюдения: каждый лист ассоциируется с одним из классов, и объекту, который распределяется в лист, присваивается соответствующая метка класса.

Под обучением дерева понимается определение его структуры, операций сравнения, пороговых величин и сравниваемых на каждом узле признаков, а также ответов в каждом листе. По своей сути, дерево может быть описано как набор операций «разрезания» признакового пространства гиперплоскостями, проходящими через пороговые значения признаков. Именно поэтому дерево решений является линейным классификатором, несмотря на его достаточно сложную организацию. Стоит отметить, что в общем случае деревья применимы и для решения задач регрессии

Алгоритмы построения деревьев решений относят к категории жадных алгоритмов. Алгоритм считается жадным, если допускает, что локально-оптимальные решения на каждом шаге (разбиения в узлах), приводят к оптимальному итоговому решению. В случае деревьев решений это означает, что если один раз был выбран атрибут, и по нему было произведено разбиение на подмножества, то алгоритм не может вернуться назад и выбрать другой атрибут, который дал бы лучшее итоговое разбиение. Поэтому на этапе построения нельзя сказать обеспечит ли выбранный атрибут, в конечном итоге, оптимальное разбиение. Более подробно про алгоритмы обучения – в материалах курса, а также в дополнительных материалах.

Одно дерево не всегда может эффективно справиться с задачей классификации или регрессии. В таком случае возможным выходом является использование ансамблей. Ансамбль – это некоторая совокупность алгоритмов, объединенных в единое целое. Каждый алгоритм имеет свою вероятность ошибки, и объединяя выходы тысячи среднеточных моделей можно добиться более точного сведенного результата «голосования», усреднив результаты.

Бэггинг (от Bagging - Bootstrap aggregation) — это один из первых и самых простых видов ансамблей. Он был придуман Ле́о Бре́йманом в 1994 году. Бэггинг основан на статистическом методе бутстрэпа, который позволяет оценивать многие статистики сложных распределений, когда выборка дробиться на множество подвыборок и на них оценивается бутстрэп статистика. Бэггинг позволяет снизить дисперсию обучаемого классификатора, уменьшая величину, на сколько ошибка будет отличаться, если обучать модель на разных наборах данных, или другими словами, предотвращает переобучение. Эффективность бэггинга достигается благодаря тому, что базовые алгоритмы, обученные по различным подвыборкам, получаются достаточно различными, и их ошибки взаимно компенсируются при голосовании, а также за счёт того, что объекты-выбросы могут не попадать в некоторые обучающие подвыборки.

Случайный лес (random forest) — бэггинг над решающими деревьями, при обучении которых для каждого разбиения признаки выбираются из некоторого случайного подмножества признаков. Для задачи классификации итоговое решение выбирается по большинству результатов, выданных классификаторами, а в задаче регрессии — по их среднему значению.

Одной из главных проблем случайного леса является его склонность к переобучению и «рваным краям» разделяющих поверхностей. Алгоритмы построения и обучения лесов постоянно совершенствуются и дополняются.

Используемые библиотеки: numpy, scikit-learn, scikit-plot, matplotlib

1. **КОД ПРОГРАММЫ. ПОЯСНЕНИЕ ЭТАПОВ**

Для начала импортируем необходимые библиотеки и функции генерации данных, разработанные в предыдущих лабораторных работах. (Рис.1) Также был создан файл estimation, который включает в себя расчёт точности, чувствительности и специфичности (Рис.2)

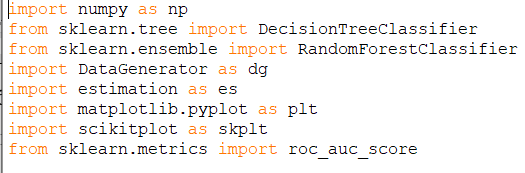


Рис. 1 — Начало работы

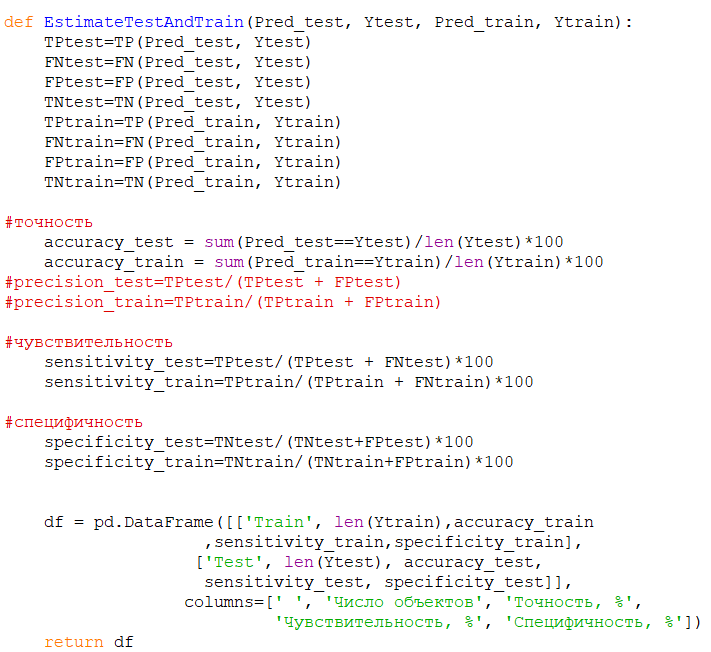
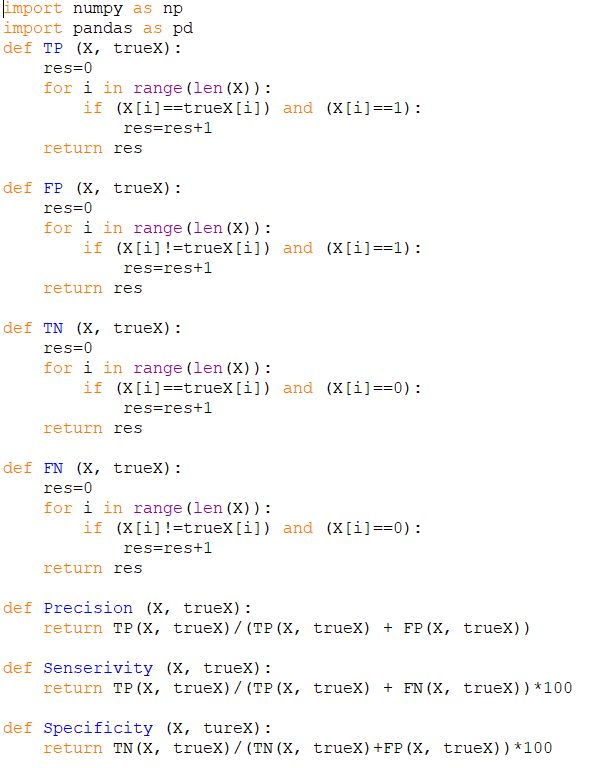


Рис. 2 — Содержание файла estimation

Используя функцию и скрипты, разработанные в лабораторной работе №1, создадим четыре переменных: Xtrain, Ytrain; Xtest, Ytest. Здесь X –объекты двух классов, параметры которых распределенны нормально, а объекты классов хорошо линейно разделимы. Y – метки класса. Для обучения и тестирования модели будут использованы разные выборки.

Вместо LogisticRegression, как в прошлой лабораторной работе, загрузим дерево решений (DecisionTreeClassifier), обучим модель на обучающей выборке, используя метод fit(). (Рис.3)

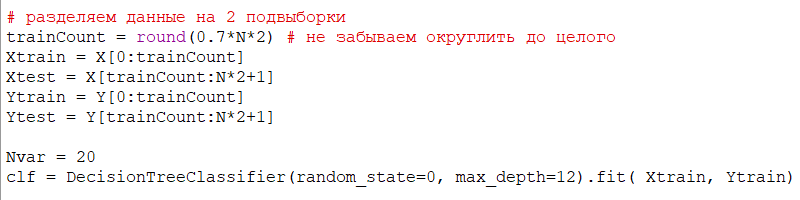


Рис. 3 — Создание переменных и обучение модели

Оценим точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках, используя метод из предыдущей лабораторной работы. (Рис.4)



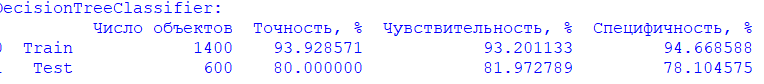


Рис. 4 — Точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках для дерева решений.

Повторим выше исполненные операции с лесом решений вместо древа решений. Можно заметить, что лес дает более точные результаты на тестовой выборке. (Рис.5)



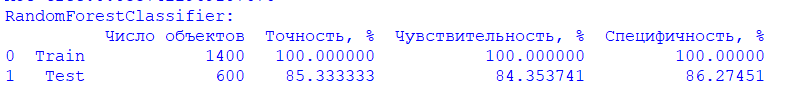


Рис. 5 — Точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках для дерева решений.

Можно заметить, что в древо решений мы передаем максимальную глубину древа. Она была найдена эмпирическим путем, и это ограничение необходимо, чтобы снизить переобучаемость модели. В лес решений мы передаем количества деревьев в лесе, которое дает наибольшее значение под кривой для тестовой выборки (Рис. 6)

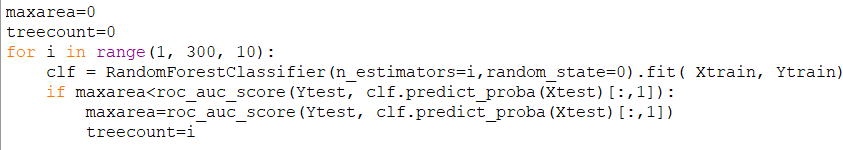


Рис. 6 — Подбор количества деревьев.

Построим ROC-кривые для дерева и для леса (Рис.8-9), используя результаты классификации тестового набора данных. Рассчитаем площадь под кривой. (Рис.8)



Рис. 7 — Расчет площади под кривой для дерева и для леса

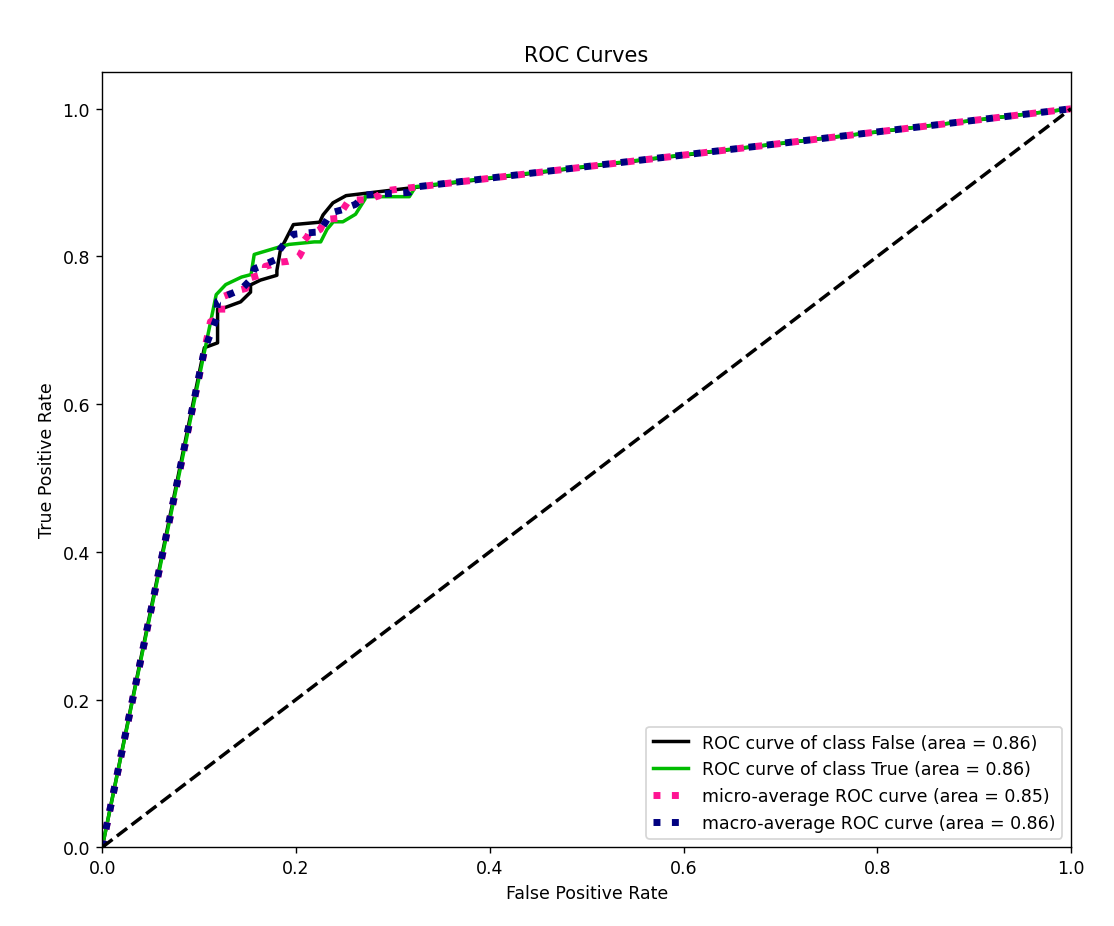


Рис. 8 — ROC-кривые для дерева

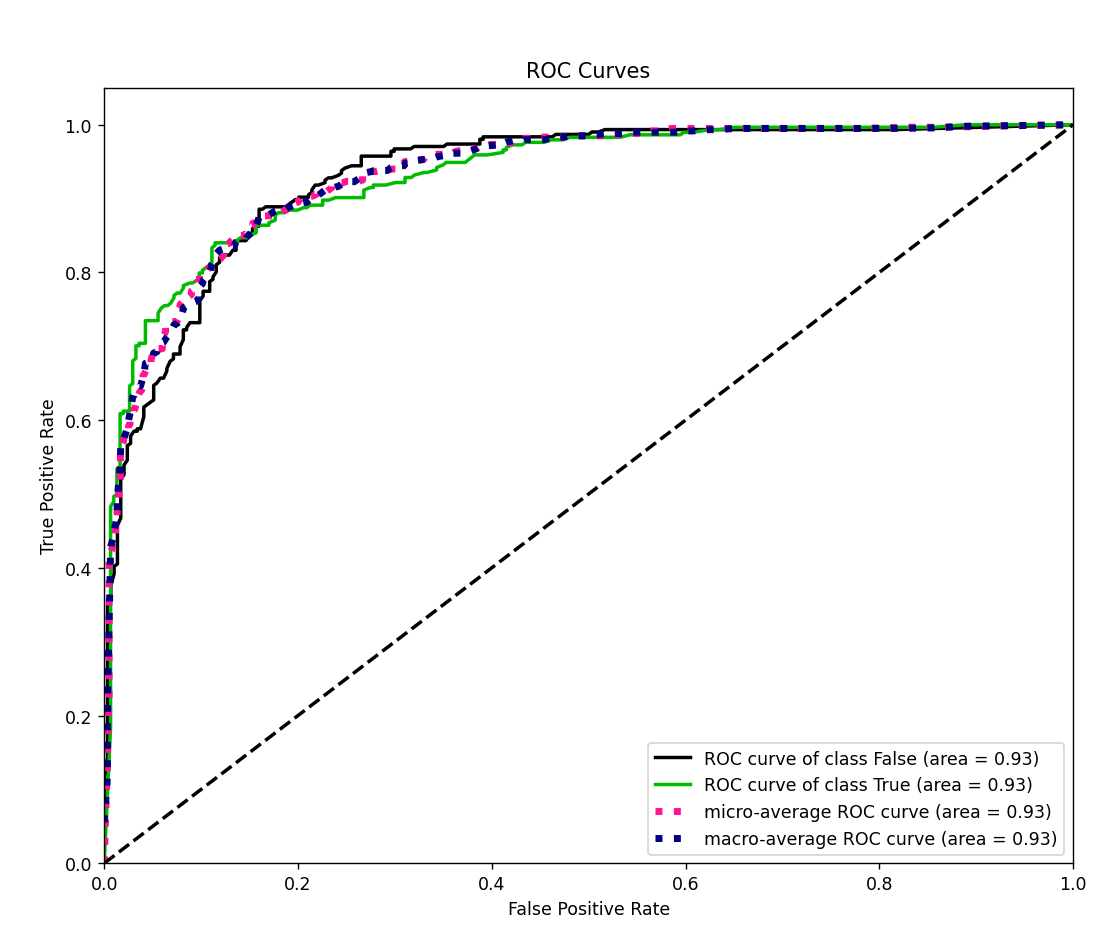


Рис. 9 — ROC-кривые для леса

Построим гистограммы распределения результатов классификации (используя вероятности) для случайного леса (для обучающей и тестовой выборок).

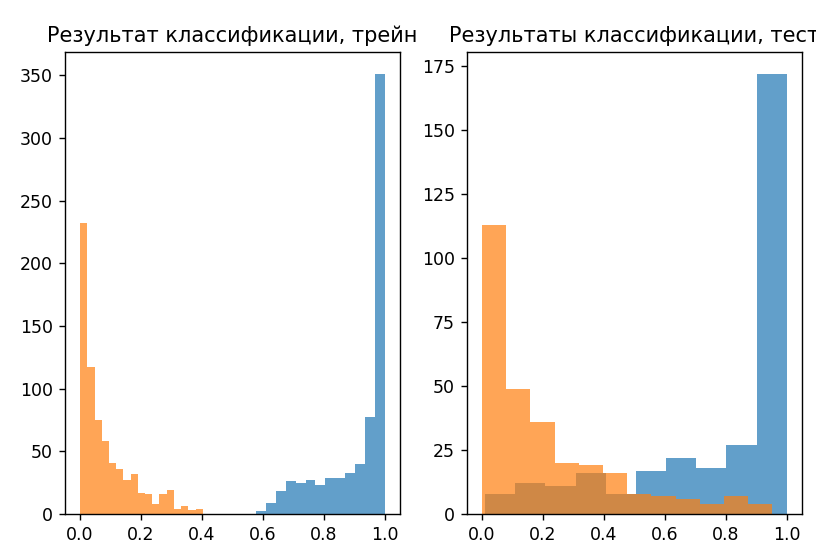
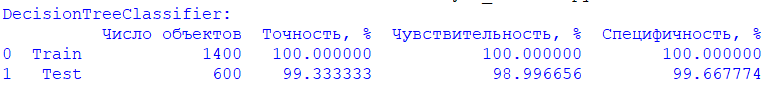


Рис. 10 — гистограммы распределения результатов классификации (используя вероятности) для случайного леса (для обучающей и тестовой выборок).

Оценим точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках для нелинейно пересекающихся классов. (Рис.11)



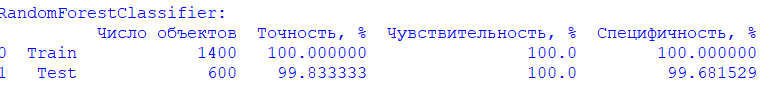
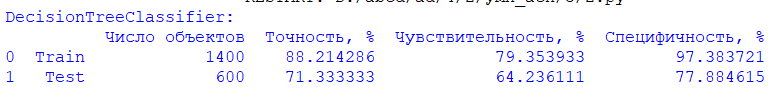


Рис. 11 — Точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках для дерева и леса решений.

Так же посмотрим на результаты распределения вероятностей для нормального распределения для плохо разделимых классов. (Рис. 12)



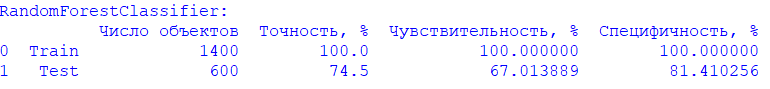
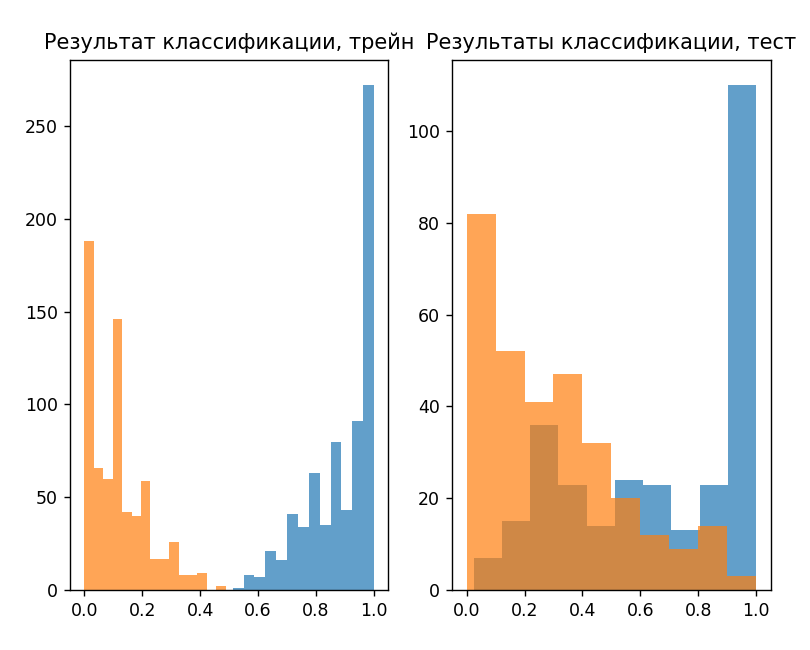
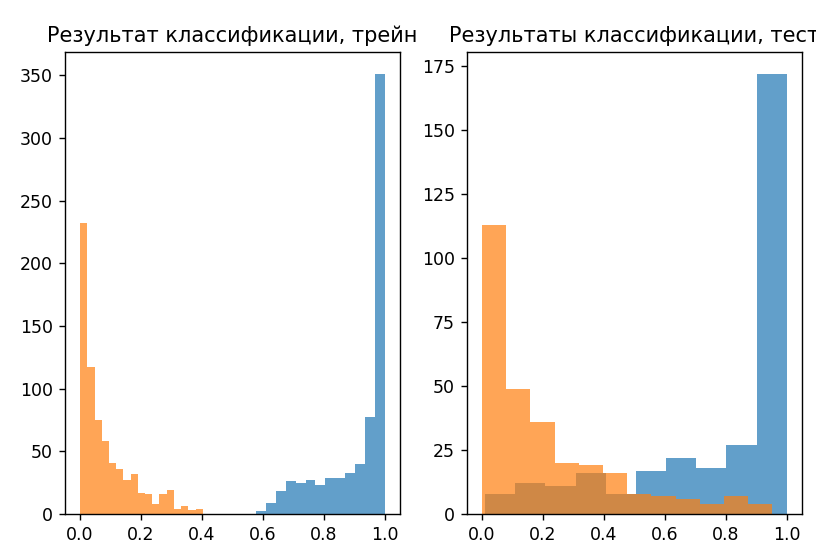


Рис. 12 — Точность, чувствительность и специфичность обученной модели на обучающей и тестовой выборках для дерева и леса решений.

Гистограммы распределения вероятности на выходе классификатора и таблицы с данными об эффективности для 3х выборок (нормальное распределение: хорошо и плохо разделимы, а также нелинейно-пересекаемые) (Рис.13)



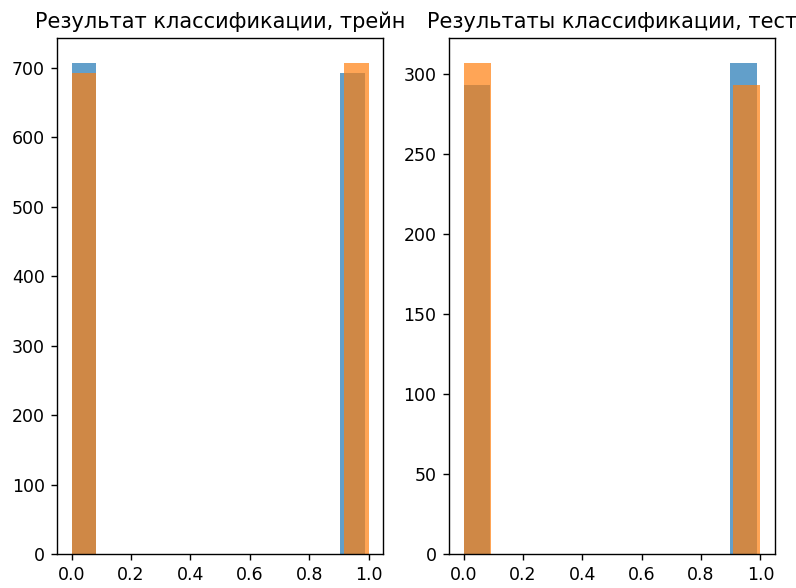


Рис. 13 — Гистограммы для трех выборок

**Выводы**

В ходе данной работе были получены навыки реализации классификатора на основе дерева и леса принятия решений и исследованы их свойства. Было выяснено, что при слишком большой глубине деревьев происходит переобучение, что было поправлено путем задания максимальной глубины дерева. Были также освежены знания по оценке классификатора без использования библиотечных формул. По трем выборкам (нормальное распределение хорошо и плохо разделимое, нелинейно-пересекаемые классы), по сравнению с результатами прошлой лабораторной работы, можно сделать вывод, что древо и лес решений справляются с задачей классификации лучше, чем классификатор на основе логистической регрессии. Особенно это можно заметить на выборке нормального распределения плохо разделенных классов. Также в данной работе были построены ROC-кривые для дерева и для леса и рассчитаны площади под этой кривой.