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

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Машинное обучение»

Тема: Исследование алгоритмов классификации

Студентка гр. 1373 _____ Новикова А.С.

Преподаватель _____ Татчина Я.А.

Санкт-Петербург

2023

Цель работы: оценить и сравнить результаты классификации, используя алгоритмы kNN и дерево решений, сравнить полученные результаты с помощью метрик качества и объяснить их.

1. Краткое описание датасета

Для этой работы был выбран другой набор данных, так как прошлый не подходил для задачи классификации.

Это набор данных признаков опухоли головного мозга. Он включает 5 признаков первого порядка и 8 признаков второго.

В датасете представлены следующие атрибуты:

- Class - целевой класс. 1 = опухоль, 0 = нет опухоли
- Mean - среднее значение (1 порядок)
- Variance - дисперсия (1 порядок)
- Standart deviation - стандартное отклонение (1 порядок)
- Entropy - энтропия (2 порядок)
- Skewness - ассиметрия (1 порядок)
- Kurtosis - эксцесс (1 порядок)
- Contrast - контраст (2 порядок)
- Energy - энергия (2 порядок)
- ASM - второй угловой момент (2 порядок)
- Homogeneity - однородность (2 порядок)
- Dissimilarity - непохожесть (2 порядок)
- Correlation - корреляция (2 порядок)
- Coarseness - грубость (2 порядок)

2. Оценить, насколько набор данных подходит для решения

В качестве целевого класса был выбран атрибут «Class», который показывает наличие опухоли головного мозга у человека. Датасет довольно сбалансирован: количество элементов, равных 1, равно 1683, а равных 0 – 2079.

3. Оценить и сравнить результаты классификации

- Дерево решений

На обучающей выборке средняя доля верных ответов составила 0. 9840513416209118. После того, как мы обучили модель, лучшее качество

составило 0.9821487579454693. Максимальная глубина = 5, макс. число признаков, которые нужно перебирать = 0.7, то есть нам необязательно перебирать все параметры, чтобы верно спрогнозировать наличие опухоли.

Получившееся дерево изображено на рис. 1.

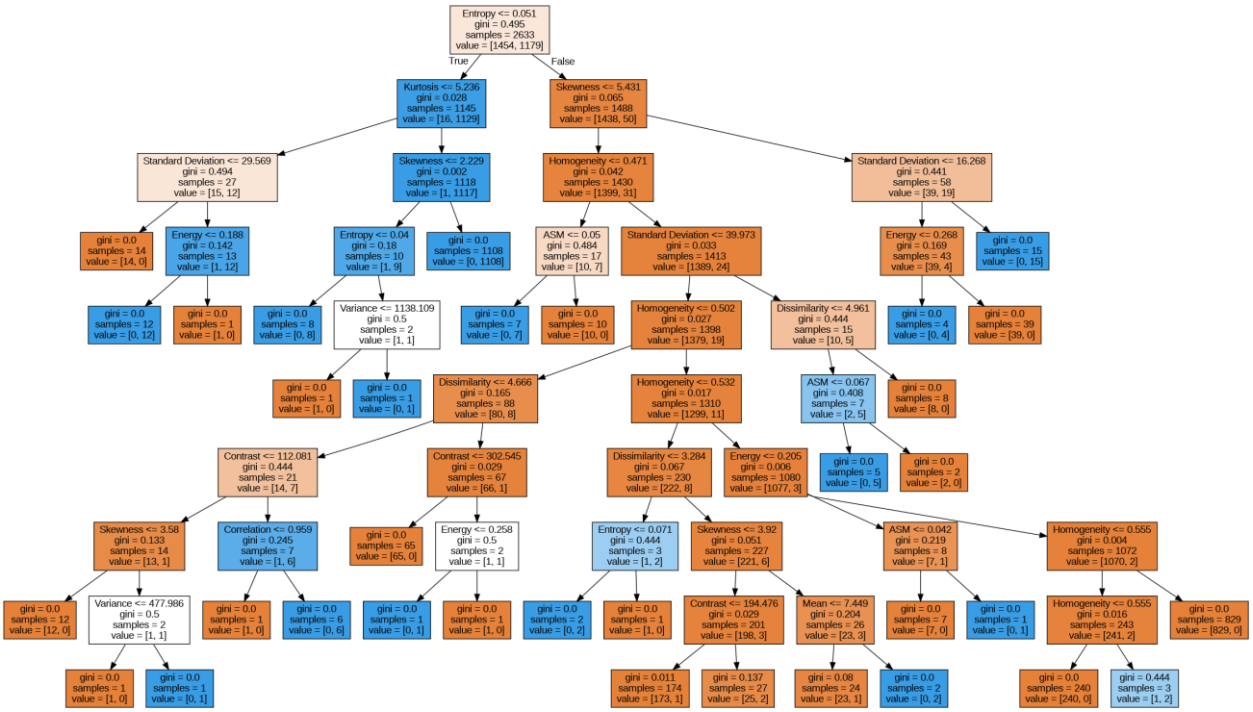

Рисунок 1. Дерево решений

- kNN

На обучающей выборке средняя доля верных ответов составила 0. 8066882634324429. После обучения лучшее качество составило 0.8188375264247731. Оптимальное число ближайших соседей = 3, то есть нам будет достаточно посмотреть на три ближайших значения, чтобы верно спрогнозировать ответ.

Во время работы на обучение дерева решений понадобилось 9.53 секунды, а kNN – 3.68 секунды. Но, несмотря на это, дерево решений даёт результат, который верен почти во всех случаях, а именно в 98%, в то время как kNN даёт верный ответ только в 82%. В нашем случае получается, что хоть и kNN обучается почти в три раза быстрее, он не даёт настолько же точные результаты, как дерево решений.

4. Сравнить полученные результаты с помощью метрик качества

- Accuracy

Дерево: 0.9813994685562445

kNN: 0.8193091231178034

Дерево предсказывает верный ответ в 98% случаев, kNN – в 82%.

- Precision

Дерево: 0.9782178217821782

kNN: 0.8177966101694916

Для дерева доля правильных ответов модели в пределах класса составила 98%, для kNN – 81%.

- Recall

Дерево: 0.9801587301587301

kNN: 0.7658730158730159

Доля предсказанных объектов, действительно относящихся к положительному классу, у дерева составила 98%, у kNN – 77%.

- F-measure

Дерево: 0.9791873141724479 – 98%

kNN: 0.790983606557377 – 79%

- ROC

ROC-кривая для дерева изображена на рис. 2, для kNN – на рис. 3.

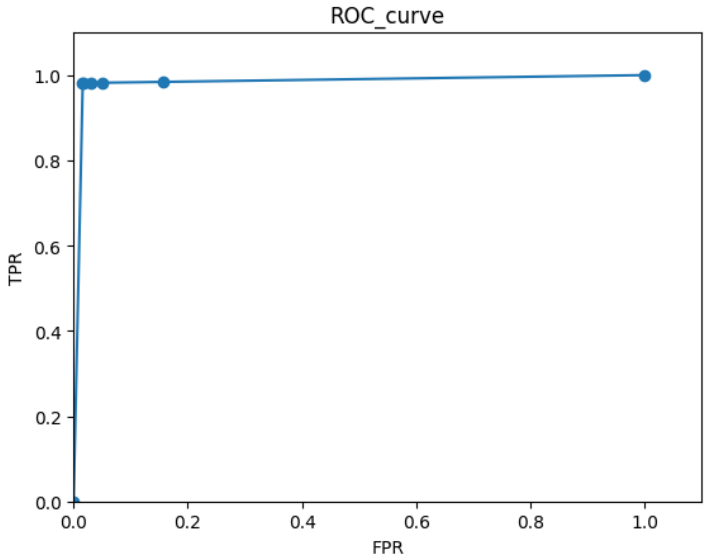

Рисунок 2. ROC-кривая дерева решений

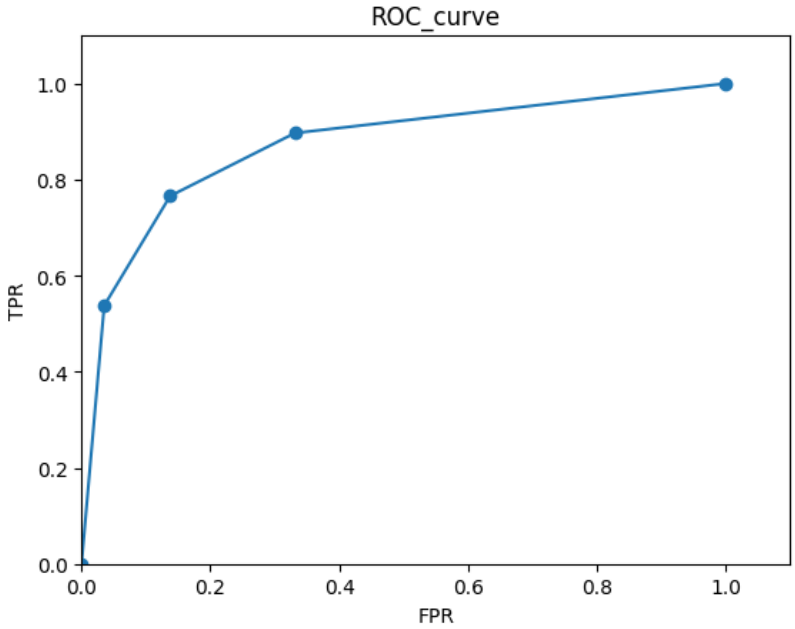

Рисунок 3. ROC-кривая kNN

На графиках мы видим, что кривая дерева выше, чем кривая kNN. Это значит, что в данном случае дерево работает лучше.

Для всех 5 метрик дерево решений показало лучшее качество, чем kNN.

5. Выводы

В ходе работы мы рассмотрели такие методы классификации, как дерево решений и kNN на датасете для выявления опухоли головного мозга. С нашей задачей по всем параметрам лучше справился метод дерева решений, так как все его оценки не падают ниже 0,97, а это значит, что метод показывает точный результат почти в 100% случаях, в то время как оценки метода kNN находятся в диапазоне от 0,78 до 0,82.