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

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №2**

**по дисциплине «Машинное обучение»**

**Тема: «Исследование алгоритмов классификации»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чесноков М. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Татчина Я. А.

Санкт-Петербург

2024

**Цель работы:** Исследовать алгоритмы классификации

**Краткое описание датасета**

Датасет о погоде. Датасет был взят с сайта [www.kaggle.com](http://www.kaggle.com/).

Почти все данные в нём числовые, только в weather содержатся строковые. В датасете представлены следующие атрибуты:

* data - дата измерений
* precipitation - выпавшие осадки в мм
* temp\_max - максимальная температура в этот день в градусах Цельсия
* temp\_min - минимальная температура в этот день в градусах Цельсия
* wind - скорость вертра в м/с
* weather - погодные условия. Целевой класс. Принимает значения 0 - "drizzle" (морось), 1 - "rain" (дождь), 2 - "sun" (солнце), 3 - "snow" (снег), 4 - "fog" (туман)

В качестве целевого класса был выбран атрибут «weather», который показывает погодные условия. Датасет довольно сбалансирован: количество элементов "rain" и "sun" намного больше, чем других.

**Классификация kNN и дерево решений**

Перед классификацией предобработаем датасет.

**kNN**:

На обучающей выборке средняя доля верных ответов составила 0.7248868778280543. После обучения лучшее качество составило 0.7272676644622346. Оптимальное число ближайших соседей = 10, то есть нам будет достаточно посмотреть на 10 ближайших значения, чтобы верно спрогнозировать ответ.

**Дерево решений**:

На обучающей выборке средняя доля верных ответов составила 0.7249564914723285. После того, как мы обучили модель, лучшее качество составило 0.8217264183780021. Максимальная глубина = 3, макс. число признаков, которые нужно перебирать = 0.5, то есть нам необязательно перебирать все параметры, чтобы верно спрогнозировать погоду.

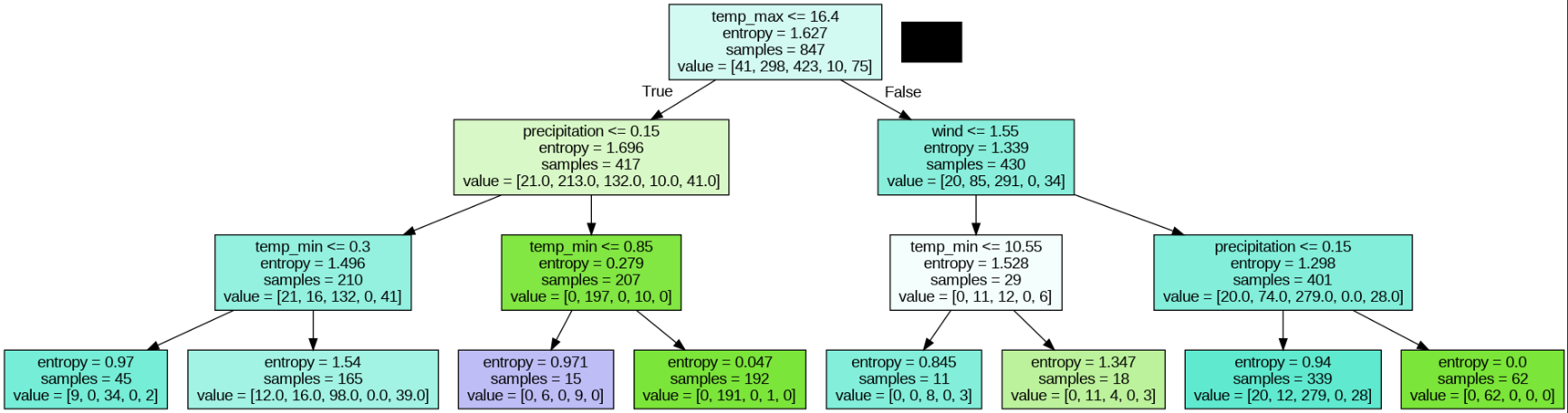


Рис. 1: Дерево решений

**Метрики оценки качества**

**Accuracy**:

kNN: 0.7554945054945055

Дерево решений: 0.8379120879120879

**Presicion**:

kNN: 0.7554945054945055

Дерево решений: 0.8379120879120879

kNN: 0.7554945054945055

Дерево решений: 0.8379120879120879

**Recall**:

kNN: 0.7554945054945055

Дерево решений: 0.8379120879120879

**F-measure**:

kNN: 0.7554945054945055

Дерево решений: 0.8379120879120879

Так как в данном датасете присутствует 5 классов, то для оценки рабочих характеристик (ROC) рассмотрим площадь под кривой:

kNN: 0.8339623361298929

Дерево решений: 0.8965237016068183

**Вывод**

В ходе данной работы я изучил метод классификации kNN и дерево решений на датасете предсказания погоды. С текущей задачей лучше справляется метод дерева решений, так как точность его результатов выше.