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

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №3**

**по дисциплине «Машинное обучение»**

**Тема: «Исследование алгоритмов классификации»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Красножён Е. Д.

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

Санкт-Петербург

2024

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

**Краткое описание датасета**

Датасет о 1215 покемонах и их характеристиках.  
В датасете представлены следующие атрибуты:

* images: Путь к изображению покемона (строковой тип. В данной работе не используется)
* index: Номер покемона в Покедексе (числовой тип. Был удалён из датасета, так как используется стандартная индексация).
* Name: Название Покемона (строковой тип).
* Type 1: Первый из типов Покемона (строковой тип).
* Type 2: Второй из типов Покемона, если присутствует (строковой тип).
* Total: Сумма всех характеристик Покемона (числовой тип).
* HP: Очки здоровья Покемона (числовой тип).
* Attack: Показатель базовой атаки Покемона для обычных атак (числовой тип).
* Defense: Базовый показатель защиты Покемона от обычных атак (числовой тип).
* SP. Atk.: Показатель базовой атаки Покемона для особых атак (числовой тип).
* SP. Def.: Базовый показатель защиты Покемона от особых атак (числовой тип).
* Speed: Определяет, какой Покемон атакует первым в раунде (числовой тип).

В качестве целевого класса был выбран атрибут «Type 1». Датасет довольно сбалансирован (Все элементы достаточно равномерно встречаются в атрибуте).

**Оценить и сравнить результаты классификации**

1. Дерево решений

На обучающей выборке средняя доля верных ответов составила примерно 0.1877. Данная модель слишком сильно настроена на обучающую выборку, поэтому необходимо переобучение. После этого, лучшее качество составило примерно 0.1686. (Максимальная глубина 4, макс. число признаков, которые нужно перебирать – 0.5)

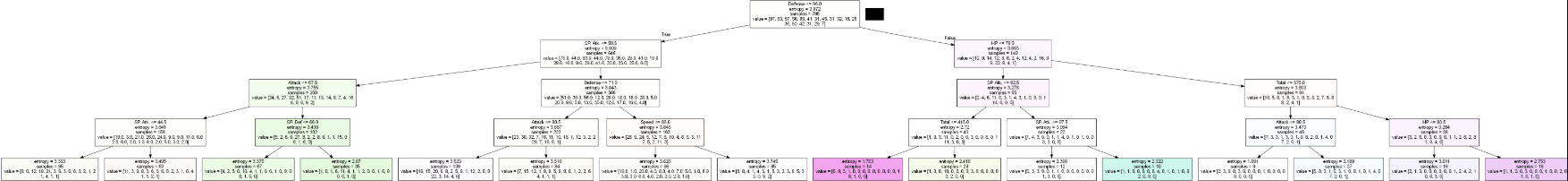


Рис. – Полученное дерево решений

1. kNN

На обучающей выборке средняя доля верных ответов составила 0.173. После обучения средняя доля верных ответов получилась 0.1836. Оптимальное число ближайших соседей = 20.

Изначально, дерево решений показывало лучший результат, но после обучения, kNN даёт более верный результат.

**Сравнить полученные результаты с помощью метрик качества**

1. Accuracy

Дерево: 0.1818

kNN: 0.2023

1. Precision

Дерево: 0.1818

kNN: 0.2023

1. Recall

Дерево: 0.1818

kNN: 0.2023

1. F-measure

Дерево: 0.1818

kNN: 0.2023

1. ROC\_AUC

Дерево: 0.58 = 58%

kNN: 0.64 = 64%

На всех метриках kNN показало более хорошее качество, чем дерево решений.

**Вывод**

В ходе данной работы были рассмотрены такие методы классификации, как дерево решений и kNN. По итогам можно сказать, что в текущем датасете практически невозможно сделать предсказание по атрибуту Type 1.В таких условиях, kNN оказался немного лучше дерева решений, опережая на сотые доли по всем метрикам качеств.