**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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чишко В. О.

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**

Данные выставления видеоиграм различных возрастных рейтингов. Представлены различные критерии, такие как: "Присутсвие алкоголя", "Присутсвие крови", "Присутсвие насилия" и т. д. Датасет взят с сайта www.kaggle.com, вот ссылка на сам датасет: <https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb>

Данные реальны. Датасет заране разделён на 2 части:

Video\_games\_esrb\_rating - обучающая и test\_esrb - отложенная.

Виды рейтингов:

* RP - Рейтинг ожидается
* EC - Для детей младшего возраста
* E - Для всех
* ET - Для всех от 10 лет и старше
* T - Подросткам
* M - Для взрослых
* A - Только для взрослых

Представлены числовые и строковые типы данных.

* title - название видеогры - (строковый)
* ESRB\_rating - возврастной рейтинг - (строковый)
* Остальные типы данных (32 столбца) сведетельствуют о наличии тех или иных критериев и принимают значения 0 и 1 - (числовой)

**Добавление нового атрибута**

В качестве целевого класса был выбран атрибут «ESRB\_rating», который показывает возрастной рейтинг видеоигры.

**Оценить и сравнить результаты классификации**

1. Дерево решений

На обучающей выборке средняя доля верных ответов составила 84.6%, но дерево получилось слишком переусложнённым. После обучения, максимальная глубина = 5, а число признаков, которые нужно перебирать = 0.7. По итогу доля верных = 80%.

1. kNN

На обучающей выборке средняя доля верных ответов составила 77.6%. После обучения средняя доля верных ответов = 81.3%. Оптимальное число ближайших соседей = 5.

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

**Сравнить полученные результаты с помощью метрик качества**

* Accuracy

Дерево: 0.76

kNN: 0.776

* Precision

Дерево:

kNN:

* Recall

Дерево: 0.76

kNN: 0.776

* F-measure

Дерево: 0.76

kNN: 0.776

* ROC\_AUC

Дерево: 0.9017666161115812 90%

kNN: 0.9199747987346704 92%

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

**Вывод**

В ходе данной работы были рассмотрены такие методы классификации, как дерево решений и kNN на датасете возрастных рейтингов. По итогу можно сказать, что оба метода оказались хороши и показывают верные результаты в большинстве случаев. kNN оказался лучше дерева решений, но оба не далеко ушли от друг друга.