**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №3**

**по дисциплине «Машинное обучение»**

Тема: Исследование алгоритмов классификации

|  |  |  |
| --- | --- | --- |
| Студент гр. |  | Лемешко А. Д. |
| Преподаватель |  | Татчина Я. А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с задачей классификации, изучение методов k-ближайших соседей и дерева решений.

**Постановка задачи.**

1. Выбрать атрибут, выступающий в роли целевого класса (метки);
2. Оценить сбалансированность классов;
3. Применить методы классификации алгоритмами kNN, Дерево решений;
4. Оценить и сравнить результаты метриками оценки качества accuracy, precision, recall, f1, ROC;
5. Проанализировать полученные результаты.

**Выполнение работы.**

Для данной работы был выбран старый датасет с информацией о студентах.

Целевой меткой для классификации возьмем Stress\_Level.

При оценке сбалансированности этой метки оказалось, что соотношение уровней стресса составляет 1029/674/297, соответственно. Было принято решение провести балансировку оверсемплингом с помощью синтетического генерирования данных методом SMOTE, а также проверить результаты для несбалансированного набора данных.

Для трёх полученных наборов данных был применён алгоритм k-ближайших соседей с выбором k методом плеча. Аналогично была обучена модель дерева решений.

Т.к. во второй работе можно было заметить очень явное распределение на категории и сильную зависимость Stress\_level от 2-3 параметров, то можно наблюдать 100% точность дерева решений при максимальной глубине 3 (при срезе глубины до 2-х, точность снижается до 0.8-0.9 в зависимости от балансировки).

При экспериментах с методом kNN можно заметить, что:

1. Четкое распределение классов привело к достаточно высокой (0.93) точности без балансировки, но для достижения такой точности пришлось рассматривать 16 ближайших соседей

2. Оба метода балансировки позволили достигнуть наилучшей точности при проверке одного ближайшего соседа, что может быть полезно для оптимизации на больших данных.

3. Наилучшая точность kNN (0.96) достигнута в третьем наборе данных, в котором метод оверсемплинга SMOTE был применен ко всем классам.

**Выводы.**

В ходе работы были изучены алгоритмы классификации KNN и дерево решений. Проведено обучение моделей, оценка и сравнение результатов.