**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. |  | Полуянов В. Н. |
| 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 набор данных оценки кредитных рисков, прошедших предварительную обработку и изучение в первой лабораторной работе.

Набор данных предназначен для изучения классификации и регрессии. Целевым классом для классификации является LoanApproved – бинарный атрибут, отвечающий на вопрос «одобрен ли кредит?».

При оценке сбалансированности этого класса оказалось, что соотношение неодобренных кредитов к одобренным составляет 15220 к 4780, соответственно. Было принято решение провести балансировку методами under- и over-sampling, 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.

Результаты обучения отображены в таблице 1.

Таблица 1. Результаты классификации.



На основе полученных результатов можно сделать выводы:

* Для данного набора данных классификация деревом решений показывает лучший результат как по сравнению с KNN, так и в общем. Итог ожидаемый, ведь деревья решений хорошо справляются с бинарными метками и являются классическим примером для решения задач одобрения кредитов.
* Результаты accuracy для несбалансированных данных хорошо отражают причину недостаточности этой метрики. Для NB KNN accuracy = 0.871, хотя более точные метрики показывают результат на 10-20% ниже. Это подчеркивает важность использования разных метрик для качественного анализа результатов модели.
* Over-sampling методом SMOTE показало наилучшие результаты, ведь изначальный объем данных не слишком велик.
* Для KNN и DT, как для зависимых от балансировки методов, любая балансировка, даже простой under-sampling, улучшает результаты.

**Выводы.**

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