**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. 2372 |  | Громова В.С. |
| 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.**

В данной лабораторной работе был использован набор данных о клиентах, содержащий атрибуты: **CustomerID**, **Gender**, **Age**, **Annual Income (k$)** и **Spending Score (1-100)**. Набор данных был обработан, были проведены следующие этапы:

1. **Обработка данных**:
   * Категориальный признак **Gender** был заменен на числовой (Male = 1, Female = 0).
   * Была проведена замена пропущенных значений: числовые атрибуты были заполнены средними значениями.
   * Для целей классификации был создан новый бинарный атрибут **SpendingClass**, который указывает на высокий или низкий уровень трат клиента (основан на значении **Spending Score (1-100)**: 1 — низкие траты, 0 — высокие траты).
2. **Анализ сбалансированности классов**:
   * После создания метки **SpendingClass** было проведено исследование баланса классов. Для данной метки данные были сбалансированы с использованием метода **over-sampling**, чтобы улучшить производительность модели.
3. **Применение алгоритмов классификации**:

На полученных данных были применены два алгоритма классификации:

* **kNN (k-Nearest Neighbors)** с выбором оптимального значения k методом плеча.
* **Decision Tree (Дерево решений)** для предсказания класса с использованием признаков.

1. **Оценка качества моделей**:

Для обеих моделей были рассчитаны основные метрики качества:

* + - **Accuracy (Точность)** — доля правильных предсказаний.
    - **Precision (Точность)** — доля верных положительных предсказаний.
    - **Recall (Полнота)** — доля реальных положительных предсказаний.
    - **F1-Score (Мера F1)** — гармоническое среднее между точностью и полнотой.
    - **ROC AUC (Площадь под кривой ROC)** — оценка качества модели на основе вероятности отнесения экземпляра к положительному классу.

Результаты обучения отображены на рисунке 1.

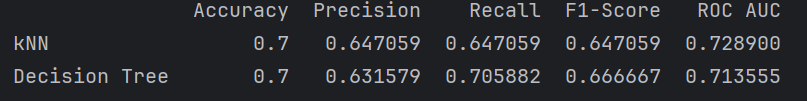
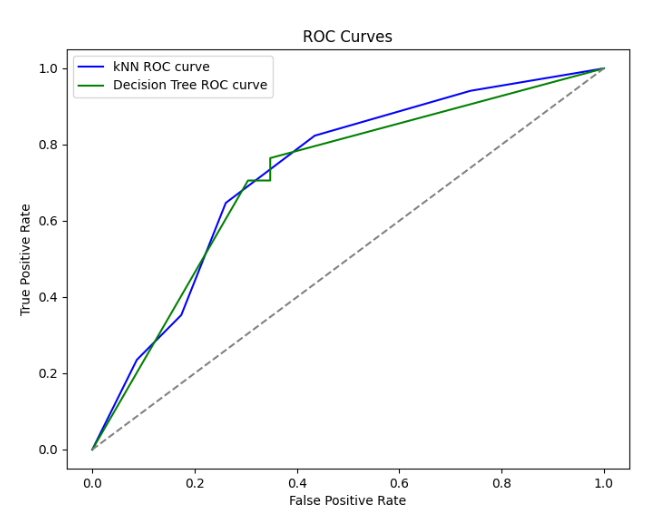


Рисунок 1. Результаты классификации.

Для каждой модели были построены соответствующие **ROC-кривые**, что позволило визуально оценить качество классификации.



1. **Результаты**:

* Модели показали хорошие результаты на сбалансированных данных, однако применение метода **over-sampling** позволило улучшить показатели **Precision** и **Recall**, что указывает на успешную балансировку классов.
* **kNN** и **Decision Tree** продемонстрировали схожие результаты, при этом **Decision Tree** показала немного лучшие результаты по метрике **F1-Score**.
* **Обе модели (KNN и Decision Tree) показывают одинаковое значение по метрике Accuracy (0.7)**, что указывает на схожую общую точность моделей. Тем не менее, различия появляются при оценке других метрик.
* **Модель kNN демонстрирует лучшие показатели по метке Precision (0.647)**, в то время как **Decision Tree** показывает меньший результат по этой метрике (0.632). Это указывает на то, что модель kNN лучше справляется с минимизацией ложноположительных классификаций.
* **Модель Decision Tree превосходит kNN по метке Recall (0.706)**, что говорит о лучшей способности модели находить все положительные примеры, но также может быть связано с возможным увеличением ложных срабатываний.
* **F1-score**, который является сбалансированной метрикой для Precision и Recall, также выше у модели Decision Tree (0.667 против 0.647 у kNN). Это подтверждает, что модель Decision Tree обеспечивает лучший баланс между точностью и полнотой, особенно в случае, когда важно избегать как ложноположительных, так и ложноотрицательных результатов.
* **ROC AUC** — метрика, которая оценивает способность модели различать классы. Здесь kNN показывает немного лучший результат (0.729), чем Decision Tree (0.714), что указывает на более высокую способность kNN различать классы в условиях различной вероятности.

**Вывод**: Несмотря на одинаковое значение Accuracy для обеих моделей, Decision Tree показывает лучшие результаты по Recall и F1-score, что делает её более сбалансированной для задачи. В то время как kNN имеет преимущество по Precision и ROC AUC, что делает её более подходящей, если важнее избегать ложных срабатываний и иметь лучшую способность различать классы.

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