Отчет по классификации гендерной принадлежности

1. Описание датасета

Датасет содержит следующие характеристики:

* Gender: Пол (целевой признак, который нужно предсказать: 0 - мужчина, 1 - женщина).

Признаки:

* Product Category: Категория продукта, который покупается.
* Age: Возраст покупателя.
* Quantity: Количество единиц продукта, купленных покупателем.
* Price per Unit: Цена за единицу продукта.

2. Описание задачи

Целью данной задачи является классификация гендерной принадлежности пользователей на основании их покупательского поведения. Для этого были применены различные модели машинного обучения, включая:

* Логистическую регрессию.
* Метод опорных векторов (SVM).
* Классификатор ближайших соседей (KNN).
* Случайный лес (Random Forest).

3. Метрики оценки

Мы используем следующие метрики для оценки моделей:

* Accuracy: Доля правильно классифицированных примеров.
* Precision: Доля истинных положительных результатов среди всех положительных классификаций (при большей важности ложноположительных результатов).
* Recall: Доля истинных положительных результатов среди всех истинных положительных примеров (при меньшей важности ложноположительных результатов).

Эти метрики помогают оценить качество моделей с разных аспектов, что особенно важно при решении задач многоклассовой классификации.

4. Результаты

После обучения всех моделей на обучающем наборе данных и предсказания на тестовой выборке, полученные результаты представлены в следующей таблице:

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Accuracy | Precision | Recall |
| Logistic Regression | 0.455 | 0.447782 | 0.455 |
| SVM | 0.505 | 0.486094 | 0.505 |
| KNN | 0.465 | 0.464728 | 0.465 |
| Random Forest | 0.465 | 0.464161 | 0.465 |

5. Выводы

В ходе анализа были протестированы четыре модели машинного обучения для классификации гендерной принадлежности пользователей на основе их покупок. Метод опорных векторов (SVM) показал наилучшие результаты среди всех используемых моделей, с точностью 0.505 и лучшим значением по метрике Precision 0.486094. Это может свидетельствовать о том, что SVM лучше справляется с разделением классов в данной задаче.

В целом, результаты моделей показывают, что задачи классификации с использованием данного датасета требуют дальнейшего анализа.