Лабораторная работа № 3

Тема: Бинарная классификация. Библиотека sklearn

1. Выберите любой доступный на просторах интернета набор данных (dataset) подходящий для **бинарной классификации**. Хорошо подходят датасеты по различным болезням, где надо определить есть болезнь/нет болезни или выдача кредитов – выдавать/не выдавать, или отток клиентов мобильных операторов – уйдут/не уйдут. Найти подходящие датасеты можно здесь:

https://www.kaggle.com/datasets?search=binary+classification

Требования к набору данных:

- минимум пять признаков, описывающих объект (можно больше);

- минимум 100 объектов для обучающего набора (можно больше);

- по набору можно оценить вероятность какого-то бинарного исхода (например, выживет/не выживет, здоровый/больной, женат/неженат, кредит выдавать/не выдавать и т.п.)

2. Проанализируйте исходные данные, при необходимости заполните пропуски или удалить не важную информацию. Категориальные признаки замените на числовые

3. Выделите из данных вектор меток У и матрицу признаков Х.

4. Разделите набор данных на обучающую и тестовую выборки.

5. На обучающей выборке получите модели дерева решений и k-ближайших соседей, рассчитайте точность моделей.

6. Подберите наилучшие параметры моделей (например, глубину для дерева решений, количество соседей для алгоритма knn)

7. Рассчитайте матрицу ошибок (confusion matrix) для каждой модели.

8. Выберите лучшую модель.

9*. Визуализируйте полученную модель дерева решений (при визуализации желательно уменьшить глубину дерева, что бы рисунок был читаемым, или сохранить в отдельный файл)

Вопросы:

1. Сформулируйте задачу классификации?
2. Что означает обучение с учителем?
3. Зачем разделять обучающую выборку?
4. Что означает переобученная модель? Как с этим бороться?
5. Что означает обобщающая способность моделей машинного обучения?
6. Объясните значения в матрице ошибок, как она рассчитывается?
7. Что показывают accuracy, precision и recall?