Обработка данных для машинного обучения

Цель работы

Познакомиться с основными приемами дескриптивного анализа данных для подготовки к машинному обучению

Задания для выполнения

1. На сайте выбрать набор данных каждому студенту индивидуально. Лучше всего выбирать датасеты в предметной области исходя из личных предпочтений. Не стоит выбирать слишком большие наборы данных (лучше не более 100 Мб).
2. В начале анализа следует привести общую характеристику набора данных: количество наблюдений, признаков, предметная область, содержание признаков, характер целевой переменной, задача обучения (регрессия, классификация, количество классов).
3. Провести в Jupyter средствами pandas, sklearn, seaborn описательный анализ выбранного набора включающего (но не ограниченного) следующие шаги:
    - исследование шкалы измерения каждого существенного признака;
    - построение эмпирического распределения каждого существенного признака;
    - анализ количества и распределения отсутствующих значений;
    - построение совместного распределения каждого признака и целевой переменной;
    - построение корреляционной матрицы.
4. По каждому пункту анализа необходимо сделать содержательный вывод.
5. Провести подготовительную обработку данных, включающую следующие шаги: 6. удаление или заполнение отсутствующих значений; 7. приведение всех признаков к бинарной либо числовой шкале; 8. удаление несущественных либо избыточных признаков; 9. другие необходимые действия в зависимости от набора данных и задачи (группировка, удаление аномалий, и др.).
6. Количественные характеристики датасета необходимо также привести после его обработки.

Контрольные вопросы

1. Какие существуют методы устранения отсутствующих значений в наборе данных? Исходя из чего нужно применять различные методы?
2. Что показывает коэффициент корреляции? В чем его ограничения?
3. Какие признаки можно считать несущественными?
4. Какие методы преобразования категориальных признаков в численные существуют?
5. Какие статистические распределения чаще всего встречаются в реальных задачах?
6. Что такое совместное и условное распределение?

Дополнительные задания

1. Использовать несколько моделей машинного обучения для решения поставленной задачи.
2. Использовать методику замеры времени обучения каждой использованной модели.
3. Сформировать сводную таблицу результатов обучения, включающую следующие данные:
    - время обучения модели;
    - accuracy;
    - precision;
    - recall;
    - f1-score.