## 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 <https://www.kaggle.com/datasets> 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:
   1. исследование шкалы измерения каждого существенного признака;
   2. построение эмпирического распределения каждого существенного признака;
   3. анализ количества и распределения отсутствующих значений;
   4. построение совместного распределения каждого признака и целевой переменной;
   5. построение корреляционной матрицы.
4. По каждому пункту анализа необходимо сделать содержательный вывод.
5. Провести подготовительную обработку данных, включающую следующие шаги:
   1. удаление или заполнение отсутствующих значений;
   2. приведение всех признаков к бинарной либо числовой шкале;
   3. удаление несущественных либо избыточных признаков;
   4. 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:
   1. время обучения модели;
   2. accuracy;
   3. precision;
   4. recall;
   5. f1-score.