Лабораторная работа №5. Предварительная обработка и очистка данных

Цель работы

Освоить методы загрузки, очистки и подготовки данных для дальнейшего анализа.

Датасет для работы

Для выполнения работы используется набор данных Titanic Dataset.

Источник: Titanic Dataset (Kaggle)

После скачивания необходимо поместить файл `Titanic-Dataset.csv` в папку проекта.

Теоретические сведения

Предварительная обработка данных является одним из важнейших этапов анализа больших данных. На данном этапе выполняются:

- загрузка данных;
- удаление пропусков;
- обработка дубликатов;
- преобразование типов данных;
- нормализация и стандартизация.

Для работы используется библиотека Pandas.

Задание

1. Загрузить набор данных из CSV-файла.
2. Определить наличие пропущенных значений.
3. Удалить или заменить пропуски.
4. Найти и удалить дубликаты.
5. Выполнить преобразование типов данных.
6. Сохранить очищенный набор данных.

Пример выполнения import

pandas as pd

```
# Загрузка данных data =
pd.read_csv("data.csv") #
Просмотр первых строк
print(data.head())

# Проверка пропусков
print(data.isnull().sum())

# Заполнение пропусков
data.fillna(data.mean(numeric_only=True), inplace=True)

# Удаление дубликатов data.drop_duplicates(inplace=True)

# Сохранение
data.to_csv("clean_data.csv", index=False)

print("Обработка завершена")
```

Контрольные вопросы

1. Что такое предварительная обработка данных?
2. Для чего выполняется удаление пропусков?
3. Что такое дубликаты?
4. Какие методы заполнения пропусков существуют?
   Для чего используется библиотека Pandas?