Практическая работа №2. Статистический анализ и визуализация данных.

Цель работы

Изучить методы статистического анализа и визуализации данных.

https://www.kaggle.com/datasets/yasserh/titanic-dataset

Задание 1: Загрузка и первичный обзор

1. Посмотрите первые 5 строк каждого файла
2. Выведите форму (количество строк и столбцов)
3. Выведите количество пропущенных значений по столбцам

Задание 2: Очистка и обработка пропусков

1. Столбец Age — заполните пропущенные значения средним возрастом или медианой
2. Столбец Embarked — заполните модой (наиболее частое значение)
3. Столбец Fare (в test.csv) — если есть пропуски — заполните средним
4. Удалите столбец Cabin или преобразуйте его (например, взять первую букву)

Задание 3: Кодирование категориальных признаков

1. Преобразуйте Sex в числовые значения (напр. female → 0, male → 1)
2. Преобразуйте Embarked (C, Q, S) в числовые категории
3. Можно создать новый признак, например, FamilySize = SibSp + Parch + 1

Задание 4: Исследовательский анализ данных (EDA)

1. Постройте гистограмму распределения возраста (Age)
2. Сравните выживаемость (Survived) по полу (Sex) — bar chart
3. Сравните выживаемость по классу билета (Pclass)
4. Проверьте распределение Fare (цена билета) — boxplot

Задание 5: Построение модели машинного обучения

1. Разделите train данные на обучающую и валидационную выборки (например, 80% / 20%) 2. Обучите простую модель, например

Logistic Regression или

RandomForestClassifier

3. Сделайте предсказания на валидационной выборке и оцените точность (accuracy)

4. Сгенерируйте предсказания на test.csv и сохраните файл для отправки

Задание 6: Улучшение модели (дополнительно)

1. Попробуйте k-fold cross-validation
2. Настройте гиперпараметры модели (grid search)
3. Попробуйте разные признаки (feature engineering) — взаимодействия, логарифмы и т.д.
4. Сравните модели: логистическая регрессия, лес, градиентный бустинг

Контрольные вопросы:

Что такое пропущенные значения?

Для чего выполняется кодирование категориальных признаков?

Что показывает гистограмма?

Что такое Accuracy?

Для чего используется train/test split?

Что такое переобучение модели?

Для чего используется cross-validation? Что такое feature engineering?