Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

“НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО”

Факультет	**Программной Инженерии и Компьютерной Техники**
Направление подготовки (специальность)	**Системное и прикладное программное обеспечение**
Дисциплина	**Системы искусственного интеллекта**

ЛАБОРАТОРНАЯ РАБОТА 3
ОТЧЕТ

Выполнил студент:	**Силинцев Владислав Витальевич (355273)**
Группа:	**P3314**
Преподаватель:	**Болдырева Елена Александровна (157150)**

г. Санкт-Петербург

2025

Содержание

ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ

- Выбор датасетов:
    - Студенты с четным порядковым номером в группе должны использовать набор данных о жилье в Калифорнии.
    - Студенты с нечетным порядковым номером в группе должны использовать про обучение студентов.
- Получите и визуализируйте (графически) статистику по датасету (включая количество, среднее значение, стандартное отклонение, минимум, максимум и различные квантили).
- Проведите предварительную обработку данных, включая обработку отсутствующих значений, кодирование категориальных признаков и нормировка.
- Разделите данные на обучающий и тестовый наборы данных.
- Реализуйте линейную регрессию с использованием метода наименьших квадратов без использования сторонних библиотек, кроме NumPy и Pandas (для использования коэффициентов использовать библиотеки тоже нельзя). Использовать минимизацию суммы квадратов разностей между фактическими и предсказанными значениями для нахождения оптимальных коэффициентов.
- Постройте **три модели** с различными наборами признаков.
- Для каждой модели проведите оценку производительности, используя метрику коэффициент детерминации, чтобы измерить, насколько хорошо модель соответствует данным.
- Сравните результаты трех моделей и сделайте выводы о том, какие признаки работают лучше всего для каждой модели.
- Бонусное задание:
    - Ввести синтетический признак при построении модели.

ОТЧЕТ О ХОДЕ ВЫПОЛНЕНИЯ

Выбор датасета

В соответствии с порядковым номером в группе (8 — чётное число) для выполнения работы был выбран набор данных California Housing, содержащий информацию о жилье в Калифорнии.

Визуализация статистики

Для первичного анализа распределения и взаимосвязей признаков были построены:

- Гистограммы распределения каждого признака
- Box-plot по каждому признаку для анализа выбросов
- Тепловая карта матрицы корреляции признаков

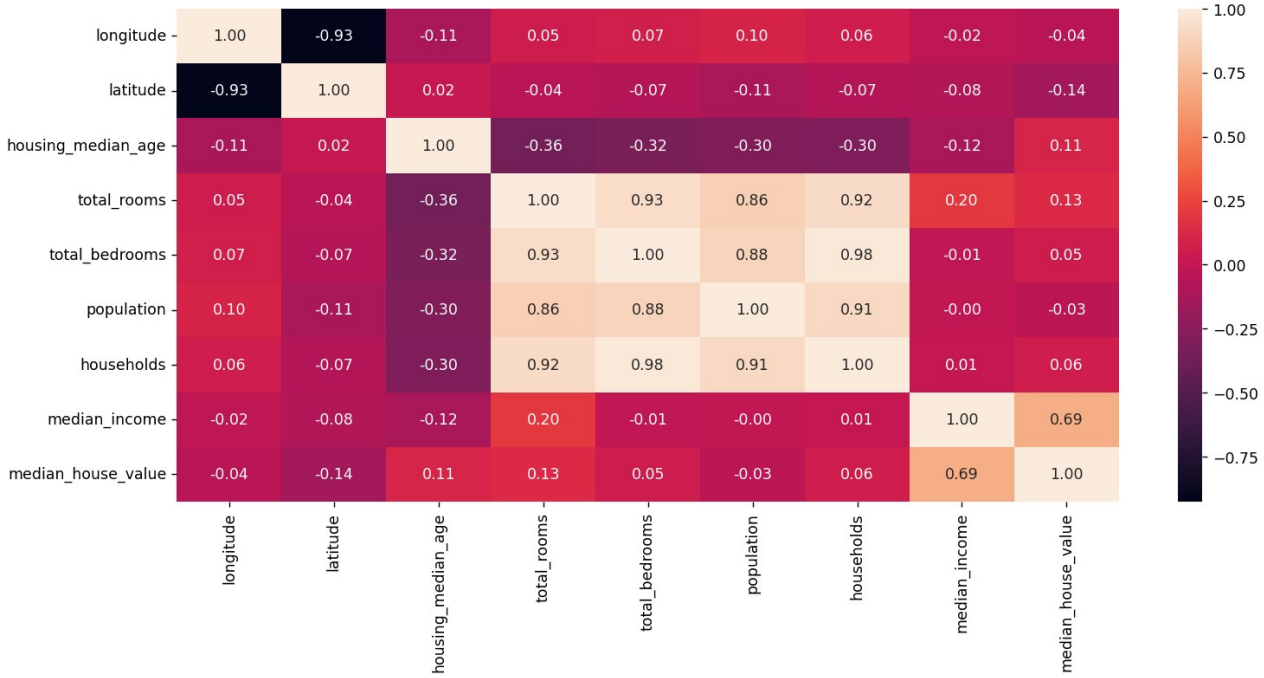

Рисунок 1 – Тепловая карта корреляции.

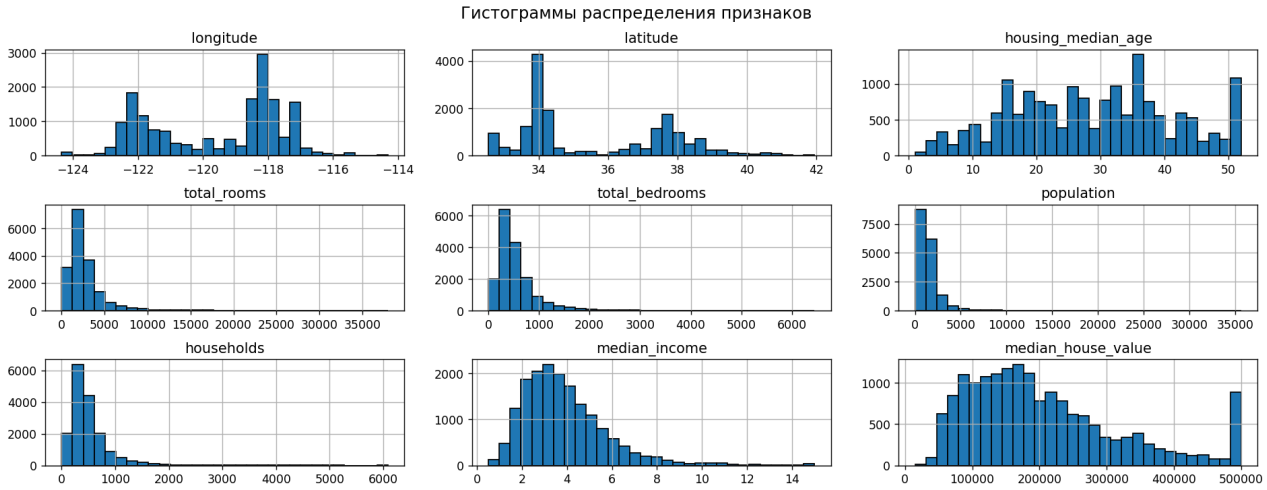

Рисунок 2 – Гистограммы распределения признаков.

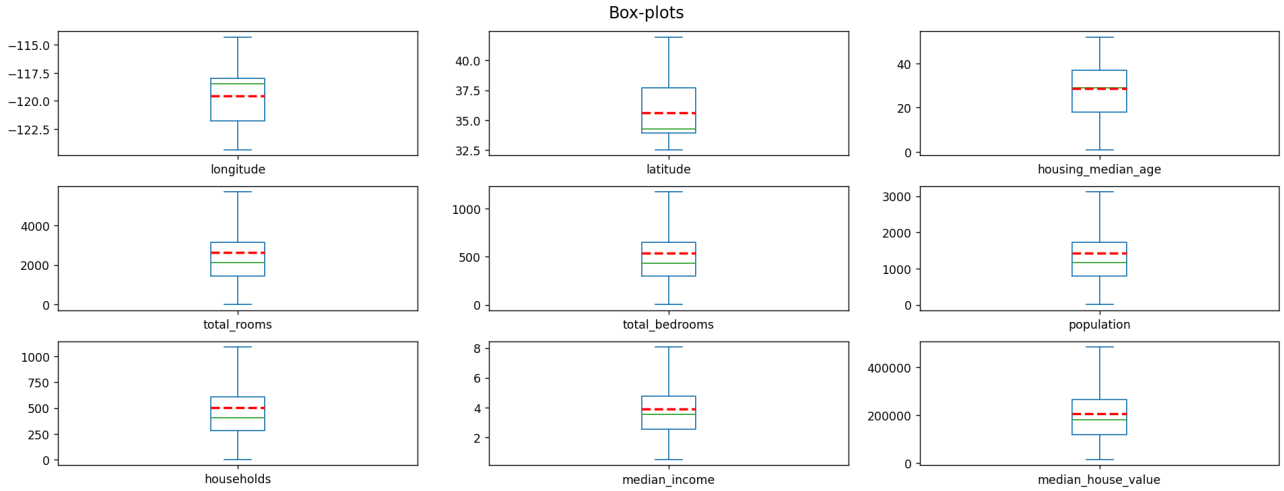

Рисунок 3 – Box-plot по каждому признаку.

Обработка данных

```
# Загрузка CSV
pd.set_option('display.max_columns', None)
df = pd.read_csv(get_resource_path("california_housing_train.csv"))

df = df.dropna()  # Убрать пустые строки
```

Разделение данных

```
target = "median_house_value"
X = df.drop(target, axis=1)  # сохранение входных признаков в переменную X
y = df[target]  # сохранение целевого признака в переменную y

# разделение на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Построение моделей

```
# Модель 1: Все признаки
model1 = LinearRegression()
model1.fit(X_train, y_train)
y_predicted1 = model1.predict(X_test)
r2_1 = r2_score(y_test, y_predicted1)

# Модель 2: Число комнат
features2 = ['total_rooms', 'total_bedrooms', 'median_income']
X_train2 = X_train[features2]
X_test2 = X_test[features2]

model2 = LinearRegression()
model2.fit(X_train2, y_train)
y_predicted2 = model2.predict(X_test2)
r2_2 = r2_score(y_test, y_predicted2)

# Модель 3: Только географические признаки
features3 = ['latitude', 'longitude']
X_train3 = X_train[features3]
X_test3 = X_test[features3]

model3 = LinearRegression()
model3.fit(X_train3, y_train)
y_predicted3 = model3.predict(X_test3)
r2_3 = r2_score(y_test, y_predicted3)
```

Оценка производительности

```
# Вывод результатов
print("=" * 100)
print("Коэффициенты детерминации:")
print(f"Модель 1 (все признаки): {r2_1: .4f}")
print(f"Модель 2 (число комнат): {r2_2: .4f}")
print(f"Модель 3 (география): {r2_3: .4f}")
print("=" * 100)

print("Важность признаков в модели 1:")
print(get_coef(X, model1))
print("=" * 100)

print("Важность признаков в модели 2:")
print(get_coef(X_test2, model2))
print("=" * 100)

print("Важность признаков в модели 3:")
print(get_coef(X_test3, model3))
print("=" * 100)
```

Функция для получения значимости коэффициентов:

```
def get_coef(x_data, model):
    """Получить коэффициенты параметров для модели"""
    return pd.DataFrame({
        'Признак': x_data.columns,
        'Коэффициент': model.coef_,
        'Абс.значение': np.abs(model.coef_)
    }).sort_values('Абс.значение', ascending=False)
```

Бонусное задание

```
# СИНТЕТИЧЕСКИЙ ПРИЗНАК:
df['bedrooms_per_room'] = df['total_bedrooms'] / df['total_rooms']
```

Разработанное приложение

Исходный код приложения: https://github.com/vvlaads/AI-systems-3.

Пример работы приложения:

```
================================================================================================
Основная статистика:
          longitude      latitude  housing_median_age   total_rooms  \
count  17000.000000  17000.000000        17000.000000  17000.000000
mean    -119.562108     35.625225           28.589353   2643.664412
std        2.005166      2.137340           12.586937   2179.947071
min     -124.350000     32.540000            1.000000      2.000000
25%     -121.790000     33.930000           18.000000   1462.000000
50%     -118.490000     34.250000           29.000000   2127.000000
75%     -118.000000     37.720000           37.000000   3151.250000
max     -114.310000     41.950000           52.000000  37937.000000

       total_bedrooms    population    households  median_income  \
count    17000.000000  17000.000000  17000.000000   17000.000000
mean       539.410824   1429.573941    501.221941       3.883578
std        421.499452   1147.852959    384.520841       1.908157
min          1.000000      3.000000      1.000000       0.499900
25%        297.000000    790.000000    282.000000       2.566375
50%        434.000000   1167.000000    409.000000       3.544600
75%        648.250000   1721.000000    605.250000       4.767000
max       6445.000000  35682.000000   6082.000000      15.000100
```

Рисунок 4.1 – Пример работы приложения.

```
       median_house_value
count        17000.000000
mean        207300.912353
std         115983.764387
min          14999.000000
25%         119400.000000
50%         180400.000000
75%         265000.000000
max         500001.000000
================================================================================================
Коэффициенты детерминации:
Модель 1 (все признаки):  0.6636
Модель 2 (число комнат):  0.5308
Модель 3 (география):  0.2423
================================================================================================
Важность признаков в модели 1:
              Признак   Коэффициент  Абс.значение
0           longitude -43465.247687  43465.247687
1            latitude -43106.304441  43106.304441
7       median_income  40194.726347  40194.726347
2  housing_median_age   1131.724382   1131.724382
4      total_bedrooms    113.707973    113.707973
6          households     45.147555     45.147555
5          population    -35.657077     35.657077
3         total_rooms     -8.843261      8.843261
================================================================================================
```

Рисунок 4.2 – Пример работы приложения.

```
Важность признаков в модели 2:
          Признак   Коэффициент  Абс.значение
2   median_income  49158.465963  49158.465963
1  total_bedrooms    166.612129    166.612129
0     total_rooms    -31.560213     31.560213
================================================================================================
Важность признаков в модели 3:
     Признак   Коэффициент  Абс.значение
1  longitude -71966.981284  71966.981284
0   latitude -70282.244356  70282.244356
================================================================================================
```

Рисунок 4.3 – Пример работы приложения.

ЗАКЛЮЧЕНИЕ

В ходе лабораторной работы было проведено исследование применения линейной регрессии с использованием языка Python и специализированных библиотек: pandas для обработки данных, matplotlib.pyplot для визуализации, numpy для вычислений и scikit-learn для построения моделей методом наименьших квадратов с оценкой их качества. Было разработано приложение, которое строит три модели линейной регрессии с разными наборами признаков, визуализирует статистические характеристики датасета, а также сравнивает эффективность моделей на основе коэффициента детерминации R^2 с анализом влияния отдельных признаков на результат предсказания. В результате проведённого анализа были определены наиболее информативные признаки для прогнозирования целевой переменной и оценено качество построенных моделей.