**Отчет по лабораторной работе: Регрессия**

**1. Введение**

Целью данной лабораторной работы является построение модели регрессии для прогнозирования цен на дома на основе характеристик, таких как площадь, количество спален, ванных комнат и других факторов. Для решения задачи были использованы различные методы регрессии, включая линейную регрессию, полиномиальную регрессию, дерево решений, случайный лес и машину опорных векторов. Каждая модель была оценена по набору метрик, таких как средняя абсолютная ошибка (MAE), среднеквадратичная ошибка (MSE), корень из MSE (RMSE), коэффициент детерминации (R²) и скорректированный коэффициент детерминации (Adjusted R²).

**2. Анализ данных**

**Этап 1: Загрузка данных**

Для анализа был использован датасет о ценах на дома, содержащий следующие столбцы:

* **price** – цена дома
* **area** – площадь в квадратных метрах
* **bedrooms** – количество спален
* **bathrooms** – количество ванных комнат
* **stories** – количество этажей
* **mainroad** – наличие главной дороги поблизости
* **guestroom** – наличие гостевой комнаты
* **basement** – наличие подвала
* **hotwaterheating** – наличие горячего водоснабжения
* **airconditioning** – наличие кондиционера
* **parking** – количество парковочных мест
* **prefarea** – наличие предпочтительного района
* **furnishingstatus** – статус меблировки

**Результат:**

* Датасет состоит из 545 записей и 13 столбцов.
* Преобразованы бинарные категориальные признаки в числовые значения для дальнейшего анализа.

Пример вывода:

# Пример вывода

Названия столбцов: ['price', 'area', 'bedrooms', 'bathrooms', 'stories', 'mainroad', 'guestroom', 'basement', 'hotwaterheating', 'airconditioning', 'parking', 'prefarea', 'furnishingstatus']

Первые 5 строк датасета:

price area bedrooms bathrooms stories mainroad guestroom basement ...

**Этап 2: Анализ данных**

* Преобразование бинарных признаков (например, наличие главной дороги, гостевой комнаты, кондиционера) в числовые значения с помощью метода .map().
* Применение One-Hot кодирования для признака **furnishingstatus** (статус меблировки), что позволило создать новые столбцы для каждого типа меблировки.

Результат:

Данные после предобработки (первые 5 строк):

price area bedrooms bathrooms stories mainroad guestroom ...

0 13300000 7420 4 2 3 1 0 ...

1 12250000 8960 4 4 4 1 0 ...

2 12250000 9960 3 2 2 1 0 ...

**3. Построение моделей**

**Этап 3: Построение моделей**

Для построения модели использовались следующие алгоритмы регрессии:

* **Линейная регрессия** — стандартный метод для линейных зависимостей.
* **Полиномиальная регрессия** (степень 2) — расширяет линейную модель, позволяя учитывать нелинейные зависимости.
* **Дерево решений** — модель, которая использует разделение данных на основе признаков.
* **Случайный лес** — ансамблевый метод, использующий несколько деревьев решений.
* **Машина опорных векторов (SVR)** — метод, использующий ядро для работы с нелинейными зависимостями.

Модели обучались на обучающей выборке и оценивались по следующим метрикам:

* **MAE (Mean Absolute Error)** — средняя абсолютная ошибка
* **MSE (Mean Squared Error)** — среднеквадратичная ошибка
* **RMSE (Root Mean Squared Error)** — корень из MSE
* **R² (Coefficient of Determination)** — коэффициент детерминации
* **Adjusted R²** — скорректированный коэффициент детерминации

**Результаты:**

Linear Regression:

MAE: 970043.4039

MSE: 1754318687330.6641

RMSE: 1324506.9601

R²: 0.6529

Adjusted R²: 0.6054

Cross-Validated R²: 0.6470

Random Forest:

MAE: 1022560.0528

MSE: 1964193399645.3335

RMSE: 1401496.8425

R²: 0.6114

Adjusted R²: 0.5582

Cross-Validated R²: 0.5947

**4. Оценка результатов**

На основании полученных метрик можно сделать следующие выводы:

* **Линейная регрессия** показывает наилучшие результаты по меткам **R²** и **Adjusted R²**.
* **Случайный лес** демонстрирует результаты немного хуже, но всё равно показывает хорошую способность моделировать данные.
* **Машина опорных векторов** показала худший результат, что связано с её сложностью в данном контексте.

**5. Визуализация**

Для визуализации зависимости между признаками и целевой переменной (ценой) была использована функция plot\_regression(), которая строит графики для каждой модели и выбранного признака.

**Результат:**

* Визуализация регрессии по признаку **area** (площадь дома) на обучающей и тестовой выборках для каждой модели (графики прилагаются).

**6. Заключение**

В ходе работы была успешно построена модель регрессии для прогнозирования цен на дома. Наилучшие результаты показала линейная регрессия, с коэффициентом детерминации **R² = 0.6529**, что означает, что модель объясняет 65% вариации данных. Несмотря на это, более сложные модели, такие как случайный лес и дерево решений, показывают более высокую способность к обобщению на тестовых данных.

**Возможные улучшения:**

* Применение более сложных методов (например, гра